Federal Aviation Administration

DOT/FAA/AM-11/8
Office of Aerospace Medicine
Washington, DC 20591

Development, Validation, and Deployment of an Occupational Test of Color Vision for Air Traffic Control Specialists

Thomas Chidester[1]
Nelda Milburn[1]
Nicholas Lomangino[2]
Nancy Baxter[1]
Stephanie Hughes[1]
L. Sarah Peterson[1]

[1]Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

[2]Office of Aerospace Medicine
Federal Aviation Administration
Washington, DC 20591

May 2011

Final Report

OK-11-0024-JAH

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications Web site:
www.faa.gov/library/reports/medical/oamtechreports

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-11/8		

4. Title and Subtitle	5. Report Date
Development, Validation, and Deployment of an Occupational Test of Color Vision for Air Traffic Control Specialists	May 2011
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Chidester T,[1] Milburn N,[1] Lomangino N,[2] Baxter N,[1] Hughes S,[1] Peterson L[1]	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
[1]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125 [2]FAA Office of Aerospace Medicine Washington, DC 20591	11. Contract or Grant No.
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved task AM-523

16. Abstract

Air Traffic Control Specialists (ATCSs) are responsible for the safe, efficient, and orderly flow of traffic in the U.S. National Airspace System. Color has become an integral element of the air traffic control environment. It is used to communicate information to ATCSs about various modes of air traffic functions including conflict alerts, aircraft control status, and weather. The Federal Air Surgeon (AAM-1) and Human Factors Research, Engineering, and Development office (AJP-61) tasked the Civil Aerospace Medical Institute (CAMI) to develop, validate, and implement an occupational test for ATCS job candidates who fail clinical instruments during the pre-employment medical examination. The Aerospace Human Factors Research Division (AAM-500) of CAMI developed the Air Traffic Color Vision Test (ATCOV) to determine whether individuals with color vision disorders (CVDs) have adequate color vision to perform critical color-related tasks involved in air traffic control. The research team conducted two studies to validate ATCOV testing. The results of Study One provided evidence of the reliability of the subtests, established performance norms for subjects with normal color vision (NCV) on each subtest, determined cut scores to apply in occupational testing, and examined the impact of testing upon a sample of CVD subjects.

The results of Study Two provided evidence of the reliability of second operational ATCOV subtests, established performance norms for NCV subjects on each subtest, determined cut scores to be applied in occupational testing, and examined the impact of testing upon a sample of CVD subjects. Color vision ability sufficient to perform duties safely remains critical to provision of air traffic services in the National Airspace System. ATCOV complies with Uniform Guidelines reporting requirements for both content and construct-oriented validity. Evidence of content validity for ATCS duties is provided through direct sampling of form and content of critical display data. Evidence of construct validity is provided by correlation with Colour Assessment and Diagnosis Test and Cone Contrast Test threshold scores, which precisely measure color vision ability. This resulted in a job sample test closely tied to critical tasks communicated using color on air traffic displays. ATCOV makes use of display formats and color chromaticities deployed for critical information on critical displays as defined by published analyses of ATCS tasks. Its items are isomorphic with datablocks and weather depictions deployed on ARTS, STARS, and DSR displays in terminal and en route facilities. Future challenges will surround the stability of color use on new systems and displays.

17. Key Words	18. Distribution Statement
Air Traffic Control, Color Vision, Personnel Selection, Medical Qualification	Document is available to the public through the Defense Technical Information Center, Ft. Belvoir, VA 22060; and the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	32	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGMENTS

This research was conducted under the Air Traffic Program Directive/Level of Effort agreement between the Human Factors Research and Engineering Group at FAA Headquarters and the Aerospace Human Factors Division of the Civil Aerospace Medical Institute.

Dr. Larry Bailey conducted internal consistency analyses in Study One and Study Two for the operational versions of ATCOV. We appreciate his knowledge of signal detection theory and methods for assessing test reliability within that approach, along with his significant efforts to complete these analyses.

CONTENTS

Development, Validation, and Deployment of an Occupational Test of Color Vision for Air Traffic Control Specialists

Air Traffic Control Specialists (ATCSs) are responsible for the safe, efficient, and orderly flow of traffic in the U.S. National Airspace System. Controllers in the cab at an airport traffic control tower (ATCT) are responsible for separating aircraft operating in close proximity to the airport and on the airport surface, including taxiways and runways. Their primary tool is direct visual surveillance of the airport area; secondary surface movement and radar displays are provided at a subset of airports. Controllers in a terminal radar approach control (TRACON) facility use radar displays to track aircraft positions in the airspace surrounding one or more airports. Controllers in air route traffic control centers (ARTCCs, or "en route centers") use radar to track and monitor aircraft positions and altitudes in flights between airports.

Color has become an integral element of the air traffic control environment. Color is used to communicate information to ATCSs about various modes of air traffic functions, including conflict alerts, aircraft control status, and weather. Color is used to draw attention to critical targets or urgent conditions, identify categories of information, and segment complex visual scenes (Xing & Schroeder, 2006a).

Color is used in a variety of ways to communicate information. For example, different colored lights are used to distinguish taxiways from runways at night. Aircraft paint schemes (known as "livery"), particularly tail signs, are used by Tower controllers to differentiate airliners; colored wing lights are used to determine aircraft orientation at night. In a TRACON or ARTCC, newer radar and other displays use color to represent weather, alerts, time, sequence, and other aircraft and airspace information.

The qualification standards for ATCS positions (Office of Personnel Management, undated) have long required controllers to have normal color vision (NCV). However, only rudimentary color schemes were utilized in early air traffic control (ATC) systems. The requirement of the color standard was successfully challenged following the Americans with Disabilities Act of 1990. As a result, the FAA was required to develop an occupational test to determine if color vision deficient (CVD) applicants had sufficient color vision to safely accomplish job duties, despite the published standard. This allowed qualification of candidates with less than normal color vision, provided they could discriminate information critical to air traffic control that is communicated using color. Job candidate color vision is assessed in a post-offer pre-employment medical examination. Clinical instruments, such as pseudoisochromatic plate (PIP) tests, are used to screen applicants during the medical examination. Candidates who are identified as having a CVD in the medical examination must be given an occupational test to determine if they can perform critical job duties.

Occupational Testing of Color Vision

The FAA's previously approved occupational color vision tests (*Flight Progress Strip* and *Aviation Lights*; Mertens, 1990; Mertens, Milburn & Collins, 1995) were developed and deployed in 1992. At that time, computer displays used in air traffic facilities were monochromatic, and color was used only for flight strip information and for aircraft, ground navigation, and obstacle lighting. With the advent of new ATC systems, these tests are no longer adequate to test candidates' ability to discriminate the range of colors used for critical information in current displays.

Recent research provides some empirical support for concerns about the impact of increasing color usage on CVD controllers who were cleared by previous occupational tests. Crutchfield and Lowe (2010) examined the frequency of operational errors among color vision deficient ATCSs who were cleared by previous occupational tests, in contrast to ATCSs with normal color vision, matched for gender, age, experience, and type of operation. Since color displays were introduced, the operational error rate among CVD controllers has increased significantly relative to their rate when only monochromatic displays were deployed. Though their current error rate was not statistically greater than that of matched NCV controllers, this may be due solely to the limited power of statistical significance tests when contrasting small groups; were the number of CVD controllers increased and the phenomenon unchanged, the result would likely be significant.

Because of these concerns about potential negative effects of increasing color usage, CVD candidates identified by clinical testing and confirmed by multiple clinical tests were temporarily assigned a pending status beginning in fiscal year 2007, awaiting validation of a new occupational test.

The Federal Air Surgeon and Human Factors Research, Engineering, and Development office tasked the Civil Aerospace Medical Institute (CAMI) to develop,

validate, and implement an occupational test for ATCS job candidates who fail clinical instruments during the pre-employment medical examination. This test would allow clearance of candidates with any type of CVD, if they possess adequate color perception to discriminate information presented in color in a manner corresponding to its display for critical ATCS job duties.

The Aerospace Human Factors Research Division of CAMI developed the Air Traffic Color Vision Test (ATCOV) to determine whether individuals with CVD have adequate color vision to perform critical color-related ATC tasks. Xing (2008a, 2008b) completed a research prototype, and the current authors were tasked to implement an operational version, compliant with all applicable standards, to serve as an occupational test for ATCS candidates who fail clinical color vision screening.

Development of the Research Prototype

Xing and her colleagues (Xing, 2006a, 2006b, 2006c, 2007a, 2007b, 2008a, 2008b; Xing & Manning, 2005; and Xing & Schroeder, 2006a, 2006b) examined all then-existing ATC displays and identified circumstances in which color use was a potential problem for CVD ATCSs. A prototype test was designed to assess whether candidates identified as having a CVD can use colors found in the ATCS workplace to detect and discriminate critical data blocks, weather blocks, and alerts as depicted on ATCS displays. Xing (2008a, 2008b) developed the prototype, completed validation research, provided norms, and recommended cut-scores for selecting future controllers with CVDs. The test was composed of four subtests: Identification, Multi-tasking, Alert Detection, and Reading Colored Text. The last was deleted before implementation because color did not make text less readable among CVD subjects. Xing designed the prototype to be a construct-based job-sample test, comprehensive of display colors sampled and conceptual of controller functions: *Comprehensive*, because color is used for a variety of information on controller displays; *Conceptual*, because we cannot expect candidates to have prior knowledge of those displays or ATC procedures. This prototype will be referred to as the "research ATCOV" for the remainder of this paper.

The FAA Office of the Chief Counsel evaluated plans to implement the research ATCOV. It was determined that implementation was appropriate, pending documentation of compliance of our validation studies with the *Uniform Guidelines on Employee Selection Procedures* ("*Uniform Guidelines*"; Equal Employment Opportunity Commission, 1978), and stipulating that further research and documentation be completed.

Assessment of the Research ATCOV

Broach (2009) completed a preliminary assessment of compliance of the research ATCOV with the *Uniform Guidelines.* Broach reviewed four reports (Xing, 2008a, 2008b; Xing & Ling, under review; Xing, Broach, Ling, Manning, & Chidester, 2009). He determined that, although the *Uniform Guidelines'* reporting requirements for content validity and construct-oriented validation studies applied to the ATCS color vision selection process, the validation study documents failed to fully comply with reporting requirements. Thus, further documentation would be required to establish compliance. While substantial evidence of construct validity was presented for color vision, sufficient evidence of content validity for ATCS duties was not. The reported studies focused on the validity of the research ATCOV with respect to clinical testing, which measures the common underlying clinical condition of CVD; this establishes construct validity. The research ATCOV measured color vision and identified deficiency with high levels of sensitivity and specificity. However, the validation study documents did not describe those ATCS job duties or tasks where color vision is critical to successful performance; nor did they clearly articulate how ATCOV subtests represented those critical job tasks that involve the use of color. The studies did not establish, as required under the *Uniform Guidelines*, the relationship of the test to the job in terms of the likelihood that a person with CVD will be able to perform the critical or important job tasks requiring color vision without hazard to others. Much of this work had, in fact, been accomplished through observations made at a variety of air traffic facilities, but the activities included in the subtests were not linked to task analyses of ATCS duties, thus failing to adequately document content validity.

Additionally, the research ATCOV was tested at the Los Angeles Pre-employment Processing Center (PEPC) during January and the New York PEPC during February of 2009. The Federal Air Surgeon approved a temporary set of criteria under which the research ATCOV could clear, but not disqualify candidates; candidates who passed the research version were cleared, but candidates who failed were returned to pending status. During PEPC testing, Aerospace Medicine personnel tested several candidates who had failed initial clinical color vision screening. Their observation of candidate testing raised further concerns about the content validity of the research ATCOV. As noted above, the research ATCOV had been designed to comprehensively represent display colors and conceptually represent ATCS activities and tasks. This resulted in the use of stimulus targets in two subtests similar in format to those used in radar displays but very different in how

colors were presented. Subtests appeared to require judgments among colors in a manner inconsistent with their use on operational displays. For example, colors used only in weather radar depiction were presented in text format, and colors used on datablocks were not identical in format (or "isomorphic") to their use on operational displays. From a perspective of ensuring candidates require no knowledge of ATC tasks, this made sense, but the approach was vulnerable to content validity challenges. In addition, we identified four problems with the subtest assessing alert detection:

1. Though the stimulus and distracter datablock colors appeared to be appropriate, their format was not isomorphic to their use on any critical display.
2. Alert display screen presentation time (one-half second) was determined by reference to cognitive psychological theory and research rather than by reference to task analysis. Xing selected presentation time to be representative of that required for recognition in a variety of visual tasks examined in the published cognitive psychology literature. But there is no task analytic requirement for controllers to detect a target in that amount of time, even though Xing (2008b) determined that 95% of persons with normal color vision could do so at the selected cut-score with one-half second presentation, and 100% could do so with one-second presentation.
3. CVD candidates who had some degree of previous ATCS training appeared to have adapted or been taught to systematically scan the display to detect alerts. This strategy may have created a disadvantage at the one-half second presentation time; one cannot methodically scan the display within that period. This heightened our concern that presentation time be closely tied to job task requirements.
4. Alerts on critical displays nearly always include some form of redundant coding, but redundant coding in a form consistent with its use on critical displays was not included in the detection test.

These operational observations reinforced the perception that the research ATCOV demonstrated construct validity for identifying CVDs among ATCS candidates, but its content validity for critical job functions was at risk. From a content validity perspective, job samples used for testing must be isomorphic with job tasks. That is, colors used in testing must be the colors deployed on critical displays to communicate information necessary to critical tasks and must be representative of their presentation on those displays. Job samples must include redundant coding as deployed on critical displays, and time presentation or limits must be justifiable by analysis of ATCS task performance. Xing (2008a) cited much of the criticality analysis provided in previous task analyses of ATCS duties but did not restrict the colors sampled in the research ATCOV to those used for critical ATCS duties. She incorporated an entire range of critical and ancillary colors into the research ATCOV. For validation based upon content sampling, compliance with *Uniform Guidelines* restricts test content to critical colors and current or near-deployment displays, presented in a manner representative of their use.

Development of the Operational ATCOV

We concluded that a new version of ATCOV would be necessary to translate the research ATCOV into an operational version. We determined that the basic format of the test could be retained, but stimulus and distracter targets must be replaced with more realistic representations of targets currently deployed on critical air traffic displays to communicate critical information. Testing would be based on the color vision demands present in the Display System Replacement (DSR), Color Automated Radar Terminal System (ARTS), Standard Terminal Automation Replacement System (STARS), and User Request Evaluation Tool (URET) displays.[1] The DSR, ARTS, and STARS are the primary radar displays used to track aircraft flow in ARTCC, TRACON, and Tower facilities. These displays show the location of aircraft, along with relevant information about each aircraft and the location of weather systems within or in the vicinity of a sector or facility. URET is a tool used for planning, primarily used in the en route facilities. It is designed to help manage aircraft flow through the sector by predicting potential conflicts with other aircraft, ground hazards, or restricted airspace based on information contained in the flight plan (Hovis & Ramaswamy, 2010a). Because selected ATCSs may be assigned to or move among terminal and en route facilities, candidates must be able to discriminate critical colors used in both the terminal and en route environments.

We accomplished a preliminary linkage of ATCS tasks to color usage by using task analyses completed by Nickels, Bobko, Blair, Sands, and Tartak (1995) and updated by the American Institutes for Research (2006a, 2006b, 2006c). Appendix A documents this linkage for critical activities and tasks. Essentially, the occupational test must ensure that for radar displays candidates can:

1. discriminate among datablocks coded in color to indicate whether they are under the control of the candidate (owned), under control of someone else (unowned), being pointed out to the candidate (pointout), or in alert status (alert; highlighted due to potential for collision, loss of communication, hijacking, or

other emergency), and from datablocks coded for non-critical purposes (such as optional highlighting)
2. discriminate each level of weather severity communicated within a display type (ARTS, STARS, DSR)
3. detect and locate datablocks in collision alert status (conflict or low altitude) within time limitations necessary to prevent collision between an owned aircraft and another aircraft, terrain, or obstacles.

Based upon this linkage, we wrote specifications of subtests for the initial version of the operational ATCOV.[2] Hereafter, these subtests will be referred to collectively as the "initial operational ATCOV." Personnel from the William J. Hughes Technical Center provided critical display documentation (Friedman-Berg, Allendoerfer, & Pai, 2008) and advice regarding how to finalize the stimulus targets. In addition, FAA Academy training facility personnel at the Mike Monroney Aeronautical Center reviewed display formats and critical information of ARTS, DSR, and URET displays and assisted us in documenting display chromaticities.

Specifications for the initial operational ATCOV included five subtests to replace three of the four subtests in the research ATCOV. The Identification subtest of the research ATCOV included colors used on several operational ATC displays to support multiple activities and tasks, and it was transformed into subtests called "Radar Identification," "Weather Identification," and "URET Identification." The research ATCOV Multitasking subtask was revised to include only critical datablock colors used on radar displays, and the research ATCOV Alert Detection subtask was modified to provide redundant coding and a task-analytic-derived test presentation time. The Reading Colored Text subtest was not incorporated. This resulted in the following subtest specifications:

1. **Radar Identification** – This subtest required discrimination among owned (white), unowned (green), pointout (yellow), and alert (red) datablocks as they are color coded on ARTS, and STARS displays. (DSR displays do not use color to communicate these functions.) Highlighted datablocks (cyan) were included as non-critical distracters that may appear on some of these displays. Examples from STARS appear in Figure 1. Examples from ARTS appear in Figure 2. Candidates were presented with eight screens of 48 datablocks each. Each datablock was representative in size, font, layout, color, and content of their implementation in the ARTS or STARS displays. Candidates searched for and mouse-clicked to select one type of datablock on each screen. There were two search screens for each datablock type and a total of 10 correct datablocks of each type over the two screens. Candidates had up to 30 seconds to complete each search screen. Scoring methodology from the Signal Detection Theory literature (Tanner & Swets, 1954) was applied to each subtest. Subtest scores were calculated for each color/type as percentage correctly identified (out of 10 possible correct datablocks for each color/type) minus percentage incorrectly identified as the search color/type (out of 86 possible incorrect selections). An overall subtest score was calculated as the average of color/type scores.
2. **Radar Multitasking** – This subtest was identical to the radar identification subtest except that a multicolor distracter screen and a simple math problem were presented between the instruction and search screen. Controlling traffic requires attention to multiple tasks including monitoring displays, making calculations, and entering data. This subtest ensured that candidates could adequately discriminate

Figure 1. STARS Datablock Format

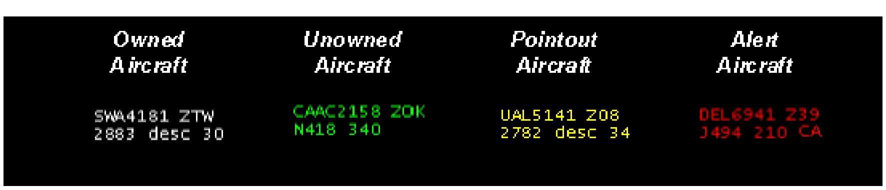

Figure 2. ARTS Datablock Format

deployed colors within a performance context rather than memorize a difference in hue or luminance for a short period of time. Subtest scores were calculated for each color/type as percentage correctly identified minus percentage incorrectly identified as the search color/type. An overall subtest score was calculated as the average of color/type scores.

3. **Alert Detection** – This subtest required candidates to quickly detect a target in alert status using color and redundant codes appearing on ARTS and STARS displays (DSR displays do not use color to indicate alert status). Redundant codes included flashing red text in positions relative to the datablock, which were specific to ARTS (to the right of the bottom line of text) and STARS (to the left and above the top line of text). Candidates were presented with a subtest screen of varying numbers of datablocks, no more than one of which was in alert status. A response screen required candidates to indicate by mouse-click whether the alert appeared on the left or right side of the screen or if no alert was present. One hundred ten subtest and response screens were presented; the first 10 were not scored. Subtest screens appeared for 2 seconds;[3] candidates must respond within 30 seconds or an incorrect response was recorded and the next subtest screen appeared. Subtest score was calculated as percentage of targets correctly identified out of 100 subtest screens.
4. **Weather Identification** – this subtest required discrimination among levels of weather intensity as they are color coded on DSR (dark blue, stippled-cyan, cyan), ARTS (dark-gray, brown, reddish-brown), and STARS (dark-gray-blue, dark-mustard) radar displays. Candidates were presented with 16 screens of 48 weatherblocks each. Each weatherblock was composed of a large outer color block (surround) and a smaller embedded color block (target). Target size was selected to subtend approximately 0.1 degree of the visual field. Mertens (1990) reported that discrimination of targets of this size was challenging along the red-green axis for NCV subjects but were representative of small targets observed on en route displays. Exemplar weatherblock targets and surrounds are depicted in Figure 3. Candidates mouse-clicked on one type of target weatherblock on each screen. There were two search screens for each weatherblock type and 10 correct weatherblocks of each type distributed between the two screens, representing all possible combinations of weather level present within display type (ARTS, STARS, or DSR). Candidates had up to 30 seconds to complete each screen. Subtest scores were calculated for each color/type as percentage correctly identified minus percentage incorrectly identified as the target search color/type. An overall subtest score was calculated as the average of color/type scores.
5. **URET Identification** – this subtest required discrimination among unalerted datalines for which the subject is responsible (described as "owned" in the subtest instructions), predicted-conflict, potential-conflict, and airspace-conflict datalines as they are color coded on URET displays: white, red, yellow, and cyan, respectively. Status-information (brown) was included as a non-critical distracter color that may appear on these displays. Candidates were presented with eight screens of 15 datalines each. Each dataline was representative in size, layout, color, and content of their implementation in URET displays. Candidates mouse-clicked on one type of dataline on each screen. There were two search screens for each dataline type and a total of 10 correct datalines of each type distributed between the two screens. Candidates had up to 30 seconds to complete each screen. Subtest scores were calculated for each color/type as percentage correctly identified minus percentage incorrectly identified as the search color/type. An overall subtest score was calculated as the average of color/type scores.

Use of High-precision Color Vision Tests in Support of Occupational Test Development

As ATCOV was developed, parallel advances in clinical color vision testing resulted in development of tests such as the Colour Assessment and Diagnosis Test (CAD; Rodriguez -Carmona, Harlow, Walker, & Barbur, 2005) and the Cone Contrast Test (CCT; Rabin, Gooch & Ivan, 2010), that precisely document the range of color perception ability of NCV and CVD individuals. In contrast to a validated job-sample test, this approach attempts to

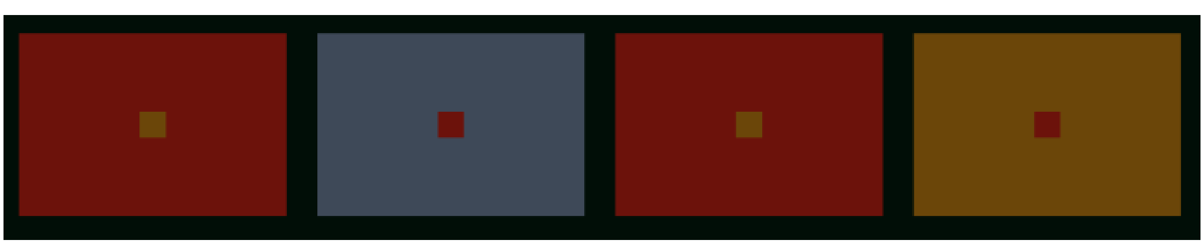

Figure 3. Weatherblock Format

more precisely measure the capabilities of a person and/or map color perception relative to a standard color space (International Commission on Illumination; CIE 1931).

The CAD is a computerized test that screens for normal color vision, quantifies loss of chromatic sensitivity, and classifies subjects by type and degree of CVD. CAD measures subjects' chromatic sensitivity threshold in 16 directions from gray (.305, .323 on the CIE 1931 color space) and scales the average threshold in the red-green and yellow-blue axes in standard normal units (SNU; standard deviations from a threshold value of zero, where 95% of the population scores less than 2 SNU). An SNU of 2 serves as the limit for a diagnosis of normal color vision (though lower values are flagged as potential deficiencies). CVD subjects scoring greater than 2 SNU have increasingly more severe deficiencies. The full-length version (definitive) CAD test takes about 15 minutes to complete; however, unlike the Nagel Anomaloscope, the diagnosis does not require an expert examiner to administer the test. The participant indicates the direction of movement of a colored target across a dynamic checkerboard background via a response pad that employs a four-alternative, forced-choice procedure with each of four buttons corresponding to the four diagonal directions of movement.

CAD threshold scores are a precise index of color sensitivity loss, and their correlation may be assessed with performance of a variety of color-based tasks. Barbur, Rodriguez -Carmona, Evans, & Milburn (2009) demonstrated, for example, that pilots with a deutan deficiency and red-green threshold scores less than 6 SNU or protan deficiencies with scores less than 12 SNU could perform as well as NCV pilots in making required performance judgments using the Precision Approach Path Indicator (PAPI) system, a highly critical flight task, while pilots with more extreme threshold scores could not. For our purposes, CAD threshold scores allow assessment of the impact of degree of CVD upon occupational test performance. Higher correlations between CAD and ATCOV scores evidence greater construct validity. Further, if CAD threshold scores reliably predict ATCOV performance, future screening by precise clinical tests might be possible. CAD became commercially available in the fall of 2009, after Study One (described below) was completed, but before ATCOV deployment and follow-on research.

The CCT is a rapid quantification test of color vision that thresholds discrimination of each cone type, providing percentage correct scores for each cone for each eye. Scores may be averaged across eyes with a passing score of 75% for each cone type. Scores can be treated as an index of sensitivity loss of each cone type. The CCT was available for use in Study Two (described below). We obtained a copy of the CCT test to assess the construct validity of ATCOV and determine whether precise scores might be used as occupationally-validated clinical tests.

Validation Research

The research team conducted two studies to validate ATCOV testing. The initial operational ATCOV was examined in Study One. Following implementation and independent assessment of *Uniform Guidelines* compliance, the test was revised into the second operational ATCOV and examined in Study Two.

Study One —Validation of the Initial Operational ATCOV

Study One was conducted to assess the reliability of the subtests, establish performance norms for NCV subjects on each subtest, determine cut scores to be applied in occupational testing, and examine the impact of testing upon a sample of CVD subjects. Cut scores were set to values described below that ensure that only CVD candidates who could discriminate critical display information communicated using colored text or weather blocks and redundant coding (text position, flashing) as well as NCV candidates would be selected. In addition, we used discriminant function, cluster, and factor analyses to gauge the dimensions underlying performance on ATCOV subtests to assess the degree to which constructs of color vision ability (normal/deficient or red-green/yellow-blue) determine subtest performance.

Method

Normative testing of NCV subjects was accomplished among 210 volunteer ATCS trainees participating in ongoing selection test validation research at CAMI. Student subjects participated during duty hours and received their regular compensation for hours worked; if they declined to participate they were given alternative tasks to fill their time. These personnel had been medically screened for normal color vision during the ATCS employment selection process. However, several individuals reported having taken secondary tests (the Aviation Lights Test and/or D-15 test) during their exam. This means that our NCV sample included some subjects with mild deficiencies identified by some, but not all, clinical tests. Subjects were screened for visual acuity of at least 20/30 in both eyes through the medical qualification process.

NCV subjects were in the early stages of their training, so they were familiar with, but not expert in air traffic tasks. Some, however, had completed Collegiate Training Initiative courses, making them more familiar than others with air traffic tasks and displays. NCV subjects were required to take the initial operational ATCOV once, but a small number (18) voluntarily took the test twice. NCV subjects were provided with practice opportunities

for each subtest; practice attempts prior to testing were not limited. Data from nine subjects were excluded from analyses because they did not respond to one or more test screens and a valid score could not be calculated, leaving a final sample of 201 NCV subjects.

Testing made use of an existing Selection Research laboratory at CAMI. The laboratory was equipped with overhead fluorescent office lights, which were illuminated during testing, producing an average of 110 cd/m^2 at the display with a chromaticity of (.4279, .4016), which is most similar to standard light source A (.4476, .4075). Office lighting was selected because candidates could be assigned to Tower (ranging from bright daylight to night exterior illumination) and TRACON facilities (dimly-illuminated windowless rooms) or to ARTCC facilities (dimly-illuminated, windowless rooms). Office lighting was a compromise among lighting conditions encountered among potential assignments.[4] All monitors in the testing laboratory were set to standard brightness (50%) and contrast (75%) settings, and variances were documented for each monitor in chromaticity of colors used in the initial operational ATCOV. Colors were set to match the RGB values used by field systems (ARTS, STARS, DSR, URET), and resulting chromaticities on the test monitors were documented.

Fifty CVD subjects who were not ATCS candidates participated to assess the relationship of ATCOV scores to CVD and predict the impact of cut-scores upon implementation. Forty-five subjects were paid volunteers recruited by a contractor from the Oklahoma City area (one of these tested as NCV on clinical tests administered in the study but was retained in the sample because he had been diagnosed previously as CVD). Five were candidates from the Great Lakes Regional PEPC who failed initial screening on the Dvorine PIP test. Volunteers were recruited through advertisements in local newspapers and online classified advertising. Volunteers were screened for at least 20/30 acuity in both eyes, with corrective lenses if required, using the Bausch and Lomb Orthorater (Bausch and Lomb, Rochester, NY). CVD was assessed by Nagel Anomaloscope (Schmidt and Haensch, Berlin, Germany). The sample included 35 protans and 13 deutans. Yellow-blue deficiency was not assessed. CVD subjects completed the ATCOV twice. CVD subjects were provided with practice opportunities for each subtest. Practice attempts prior to testing were not limited by the experimenters; subjects could access practice subtest trials until they were comfortable taking each subtest.

Prior approval for all procedures and use of human subjects was obtained from the FAA Institutional Review Board. Informed consent was obtained prior to participation and subjects were free to withdraw from the project without consequence at any time.

Results

Reliability analysis. Internal consistency of ATCOV subtests was assessed for each subject's first attempt using Chronbach's Alpha (*a*), separately among NCV and CVD subjects. Test-retest reliability was calculated only for CVD subjects because NCV subjects were not required to retest. Alpha values appear in Table 1.

Internal consistency values were acceptable for all subtests. Test-retest values among CVD subjects were excellent for Radar Identification, Radar Multitasking, and Alert Detection, and in the low acceptable range for Weather Identification and URET Identification.

Alpha values were also calculated for each individual datablock, dataline, or weatherblock type/color within each subtest. Analyses indicated that all values were acceptable among NCV subjects, with the lowest alpha value being .77 for level 3 (cyan) weatherblocks on the Weather Identification subtest. Among CVD subjects, low alpha values were obtained for owned (white) targets on Radar Identification (*a* = .58), level 2 DSR (cyan stippled) weatherblocks on Weather Identification (*a* = .56), and datalines in potential conflict (yellow) on URET Identification (a = .57). Lower alphas among CVD subjects are not surprising for these particular colors; for example, white is easily confused with cyan for individuals with a red-green deficiency.

Normal color vision (NCV) subjects. Distributions were as expected for each subtest, with scores concentrated at the upper range and tailing off sharply towards

Table 1. Internal Consistency Values for ATCOV Subtests

Subtest	*a* - NCV	*a* - CVD	*Test-retest CVD*
Radar Identification	.94	.85	.92
Radar Multitasking	.88	.94	.87
Alert Detection	.99	.95	.94
Weather Identification	.97	.84	.78
URET Identification	.95	.83	.72

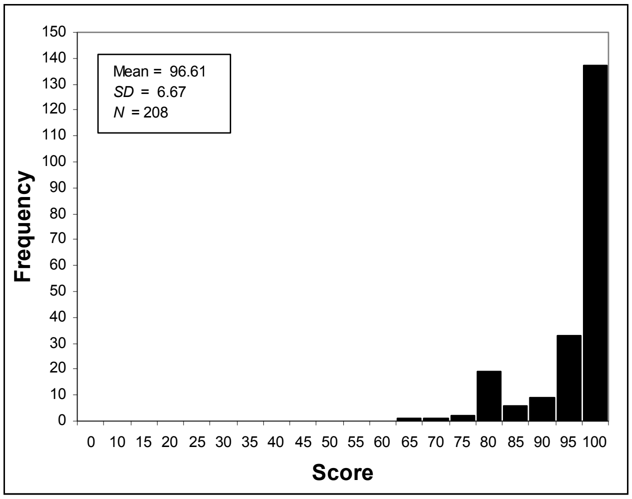

Figure 4a. Radar Identification Scores Among NCV Subjects

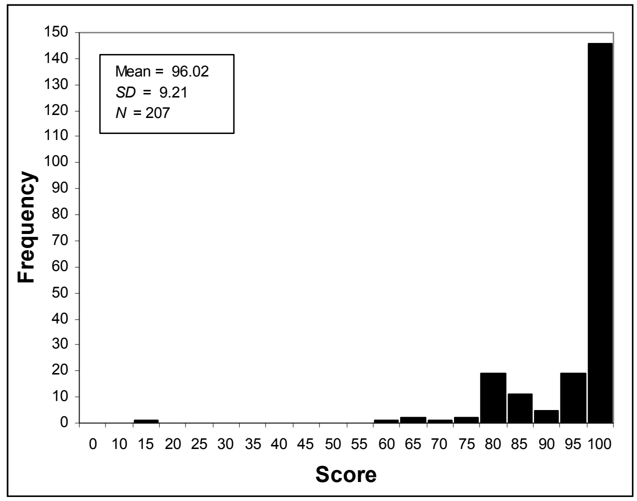

Figure 4b. Radar Multitasking Scores Among NCV Subjects

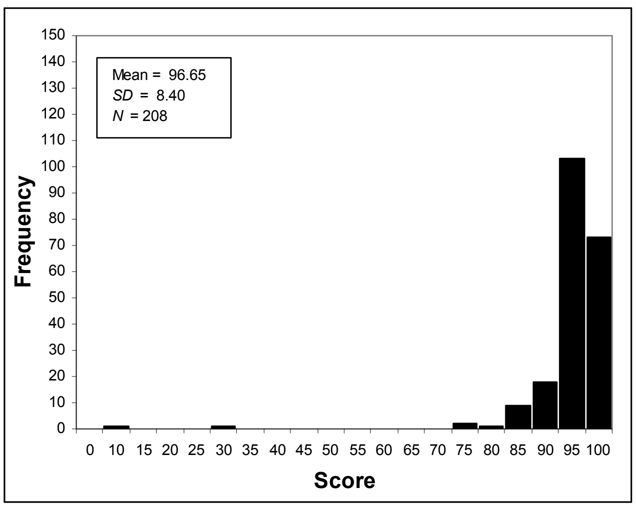

Figure 4c. Alert Detection Scores Among NCV Subjects

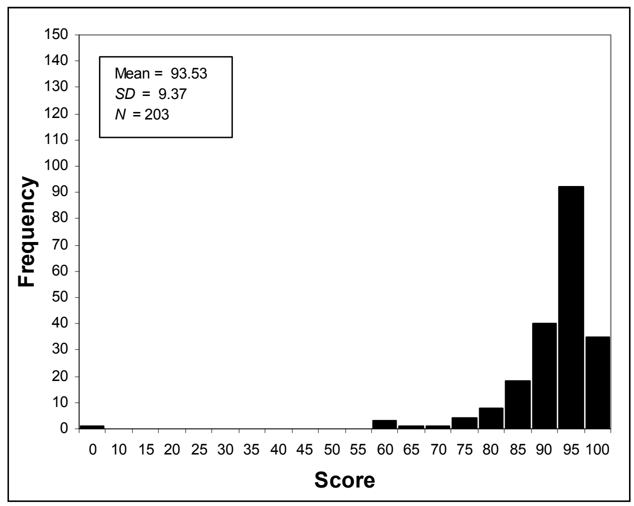

Figure 4d. Weather Identification Scores Among NCV Subjects

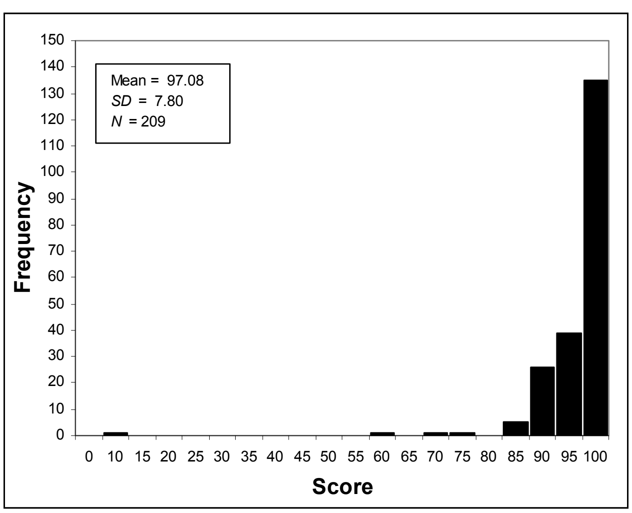

Figure 4e. URET Identification Scores Among NCV Subjects

lower scores (negatively skewed), as shown is Figures 4a through 4e.

Mean, median, standard deviation, and fifth percentile score for each subtest is documented in Table 2.

Distributions were similar to those reported for the research ATCOV. Completion time for each screen was captured for each subject for further contrast with data from CVD subjects below.

Cut-scores for each subtest were set at the fifth percentile of the 201 NCV subjects. As previous research has shown that 95% of the population perceives color in the same way (e.g., Milburn, 2004), this cutoff ensured that a candidate with a CVD could discriminate critical display information communicated using color and redundant coding as well as the NCV population, while ensuring that candidates were neither advantaged nor penalized by any residual color vision differences among the NCV

Table 2. Descriptive Statistics for ATCOV Subtests Among NCV Subjects

Subtest	Mean	Median	Std. Dev.	Fifth Percentile
Radar Identification	96.61	100.00	6.66	80
Radar Multitasking	96.02	100.00	9.21	80
Alert Detection	96.65	99.00	8.40	89
Wx. Identification	93.52	96.25	9.37	80
URET Identification	97.08	100.00	7.80	90

Table 3. Descriptive Statistics for ATCOV Subtests Among CVD Subjects

Subtest	Mean	Median	Std. Dev.	Pass 1st Attempt	Pass 2nd Attempt
Radar Identification	90.61	97.35	12.03	82%	82%
Radar Multitasking	87.78	97.62	18.62	74%	78%
Alert Detection	74.10	74.50	19.28	28%	39%
Wx. Identification	81.69	84.04	15.03	66%	66%
URET Identification	90.64	94.16	11.28	66%	78%

sample. Passing scores for the Radar Identification, Radar Multitasking, and Weather Identification subtests were set at 80. Passing score for the Alert Detection subtest was set at 89. Passing score for URET Identification was set at 90. Scores were rounded to the closest integer when compared to the cut-score. Subjects were credited with passing the initial operational ATCOV only if they passed all subtests. Applying these criteria to the NCV sample, 13.9% failed at least one test on their first attempt. Retesting was not required of these participants. However, of the 18 subjects who took the test twice, three (16.7%) failed at least one test on the first attempt, but none failed the retest. One subject failed a retest after passing an initial test.

Color vision deficient (CVD) subjects. The distributions for each subtest among CVD subjects were consistent with findings from the research ATCOV (Xing et al., 2009). On average, CVD subjects did not score as well, but a significant proportion could discriminate critical datablocks, datalines, and weatherblocks as well as NCV subjects could. Comparison to Figures 4a through 4e above reveals the distributions to be flatter (and less negatively skewed) among CVD subjects. Fewer CVD subjects scored in the 90% range and substantial numbers obtained scores in the 70% to 80% range, as shown in Figures 5a to 5e.

Mean, median, standard deviation, and passing rates on the first and second attempt for each subtest is documented in Table 3.

Overall, 22% of CVD subjects passed all subtests on their first attempt; 32% passed all subtests after two opportunities. Alert Detection was the most difficult subtest. This was expected, given the criticality of alerts and justifiable limitations on presentation time.

Analyses contrasting subtest performance of NCV and CVD subjects. Compared to NCV subjects, CVD subjects scored significantly lower on all subtests on their first attempt (Second attempt scores cannot be compared in this sample because NCV subjects were asked to complete only a single attempt). Means, effect size, and statistical significance probabilities are displayed in Table 4.

Cohen's *d* is a measure of effect size, indicating the difference between the means in pooled standard deviation units. Effects greater than .8 are considered large, .2 is considered small, and .5 moderate. A Kappa of .50 was obtained for color vision (CV) status and pass-fail results after one attempt.

Additionally, response times were significantly higher among CVD subjects, meaning they took longer to respond to or complete each screen across tests. Table 5 compares average response times for each screen, effect size, and probability level for each subtest.

Differences in response time could have practical implications for performance of ATCS duties. Since CVD subjects required more time on average to scan the display, classify the types of traffic they would be controlling or monitoring, and respond to alerts, they might be expected

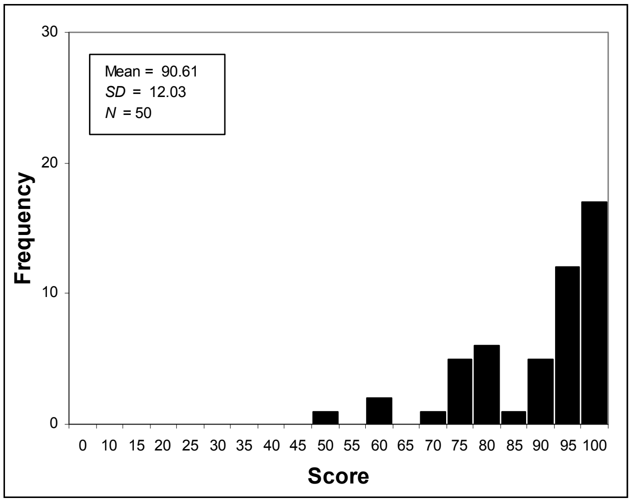

Figure 5a. Radar Identification Scores Among CVD Subjects

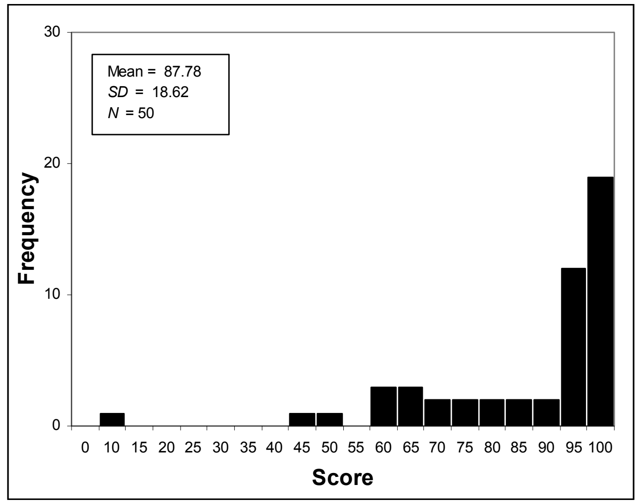

Figure 5b. Radar Multitasking Scores Among CVD Subjects

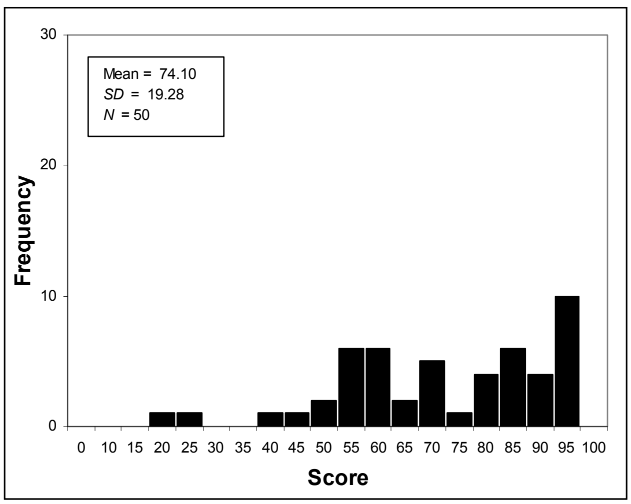

Figure 5c. Alert Detection Scores Among CVD Subjects

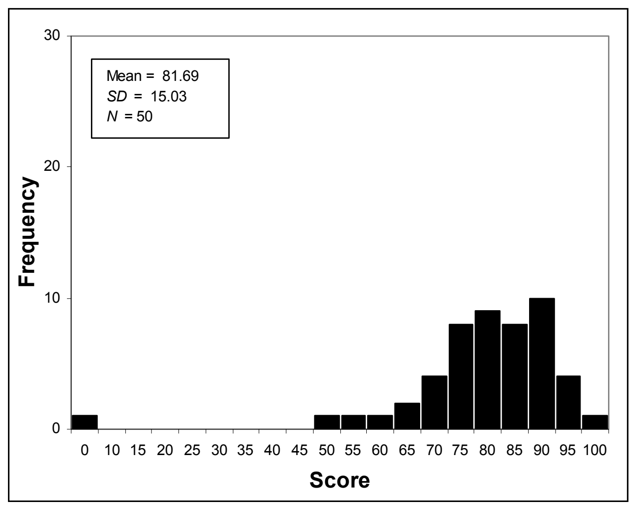

Figure 5d. Weather Identification Scores Among CVD Subjects

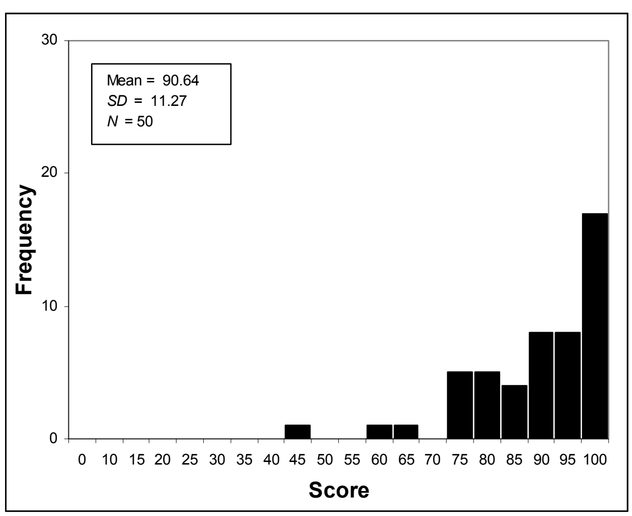

Figure 5e. URET Identification Scores Among CVD Subjects

Table 4. Comparison of Subtest Scores by NCV and CVD Subjects

Subtest	NCV	CVD	d	p
Radar Identification	96.61	90.61	.78	<.001
Radar Multitasking	96.02	87.78	.75	<.001
Alert Detection	96.65	74.10	2.14	<.001
Wx. Identification	93.52	81.69	1.13	<.001
URET Identification	97.08	90.64	.76	<.001

Table 5. Comparison of Response Time (in seconds) of NCV and CVD Subjects

Subtest	NCV	CVD	d	p
Radar Identification	9.22	20.18	1.92	<.001
Radar Multitasking	9.42	18.75	1.87	<.001
Alert Detection	6.26	9.37	1.47	<.001
Wx. Identification	21.66	24.49	.50	<.001
URET Identification	11.20	17.19	1.35	<.001

to control traffic less efficiently than NCV subjects. However, except for Alert Detection, where presentation time is set to the maximum available to enable prevention of collision, there is no defined operational requirement to respond within a specified amount of time that would justify a reaction time cut score. So, no subtest reaction time cut scores were set.

Analyses examining construct validity. If the subtests of ATCOV tap one or more underlying constructs associated with color vision ability or deficiency, we would expect subtest scores to reliably discriminate NCV from CVD subjects, that grouping subjects by similarity of subtest scores would tend to separate NCV and CVD subjects, and that the matrix of correlations among subtests could be reduced to a single or small number of factors accounting for variance in scores. We completed three analyses combining the NCV and CVD samples that are relevant to construct validity:

1. We contrasted the groups using discriminant analysis of ATCOV subtest scores. Discriminant analysis calculates an optimal weighting of predictor scores that would predict group membership and documents the classification accuracy of the solution. The discriminant function gave greatest weight to Alert Detection (.87) and about equal weight to the remaining tests (.14 to .27, all significant at the $p<.01$ level) and correctly classified 90% of subjects by NCV versus CVD status. If we had scored the subjects by this function, 99% of NCV subjects and 46% of CVD subjects would be classified in one group, and 1% of NCV and 54% of CVD subjects would be classified in a second group. Discriminant function classifications produced Kappa values of .76 for CV status and .60 for passing or failing the initial operational ATCOV.
2. We cluster-analyzed the subjects, ignoring NCV versus CVD group membership. "Cluster analysis" groups subjects by their multivariate similarity (statistical closeness) to each other. A two-cluster solution grouped 58% of CVD subjects with 99% of the NCV subjects in one group and 1% of NCV subjects with 42% of CVD subjects in the second group. Cluster membership produced Kappa values of .63 for CV status and .57 for passing or failing the initial operational ATCOV. Cross tabulation of discriminant function classifications with cluster membership resulted in a Kappa of .80.
3. We applied factor analysis (Principal Component extraction with Varimax rotation) to the five subtests to assess the dimensionality of constructs underlying performance on the Operational ATCOV. Analysis of Eigenvalues and alternative factor extraction quantities suggested that a single factor accounted for response variance (52%), presumably degree of color perception ability or deficiency. A single-factor solution weighted Radar Identification (.63) and Alert Detection (.64) highest, and Radar Multitasking (.49), Weather Identification (.41), and URET Identification (.45) approximately equally. Forced 2 and 3 factor solutions accounted for additional variance (71% and 83%, respectively) but did not reveal meaningful differences in constructs. A two-factor solution grouped Radar Identification and Radar Multitasking on one factor, Alert Detection and Weather Identification on separate factors but equally weighted URET Identification on each factor. A three-factor solution was similar but broke out

URET Identification as a separate construct. Taken together, these analyses suggest that color vision ability accounts for much performance variance among subjects completing the ATCOV. Subtest scores reliably discriminate between NCV and CVD subjects in a manner that approximates the distribution of passing and failing the test at the selected cut scores. Grouping subjects by similarity of subtest scores groups those who pass and those who fail into separate groups. Therefore, we concluded that a single factor accounts for the majority of variance in ATCOV scores.

Our sole concern with these results is that our content validity-driven scoring gave greater weight to Alert Detection and Weather Identification than we would apply if we were seeking only to measure the generic construct of color vision. This result was appropriate to a job-sample test, however, for two reasons:

1. We are sampling the critical job tasks and activities as defined by task analyses of ATCS positions. Timely responses to alerts indicating potential collision with other aircraft or terrain are critical. Separating aircraft from hazardous weather depicted on radar is also critical. Thus, empirically greater emphasis on Alert Detection and Weather Identification subtests over datablock and dataline identification is appropriate to the demands of the job.
2. Color vision deficiencies themselves are multi-dimensional and unequally distributed across dimensions (generally described along red-green and yellow-blue dimensions). Xing (2006b) showed that use of colors in air traffic displays exploits cultural meanings associated with color, (e.g., red for alerting, yellow for attention, etc.), causing some types and degrees of perceptual deficiencies to have a greater impact on job performance.

Discussion

The results of Study One provided evidence of the reliability of the subtests, established performance norms for NCV subjects on each subtest, determined cut scores to apply in occupational testing, and examined the impact of testing upon a sample of CVD subjects. In general, the subtests were internally consistent among both NCV and CVD subjects. Scores were stable among CVD subjects; stability was not assessed among NCV subjects. Cut scores were set to ensure that CVD candidates who passed the test could discriminate critical activities and tasks communicated using color as well as NCV candidates. Subtest scores separated NCV from CVD subjects on average, but identified fairly substantial numbers of CVD subjects who could discriminate critical activities and tasks communicated using color. From that perspective, the subtests functioned as desired and expected.

Discriminant, cluster, and factor analyses suggested that ATCOV subtests are correlates of color vision ability. They separate NCV from CVD subjects on average, but substantial numbers of CVD subjects are able to pass all subtests by discriminating among critical data. Subtest scores themselves appear to be tied to a single underlying construct, presumably color vision ability versus deficiency. This provides mixed evidence of construct validity for the ATCOV. To the good, ability versus deficiency classification is well-correlated with ATCOV scores. A weakness for construct validity in this study, however, is that precise measures of color vision deficiency were not collected for all subjects. We accepted that trainee ATCS subjects had been screened for normal color vision and confirmed CVD by Nagel Anomaloscope among CVD subjects but could not assess correlation between degree of CVD and ATCOV subtest scores. In addition, while color vision varies along red-green and yellow-blue dimensions, we did not see good evidence of both dimensions in the factor analysis data. We cannot determine whether this was due to the statistical predominance of red-green deficiencies in the CVD population or to potential deficiencies in the ATCOV. These weaknesses were corrected by collecting CAD and CCT data in Study Two (below).

Field Implementation of the Initial Operational ATCOV

After reviewing the results of Study One, the Federal Air Surgeon directed deployment of the initial operational ATCOV at nine regional flight surgeon (RFS) and Medical Field Offices (MFOs). Forty-three ATCS candidates who had failed clinical screening were offered and accepted an opportunity to complete occupational testing. In addition, this group was asked to take the CAD test for research purposes. These data were collected to assess whether a precise CVD diagnosis and threshold could account for outcomes of occupational testing. Twenty-two (51%) CVD controller candidates agreed to take the CAD. All obtained CAD diagnoses indicating abnormal red-green color perception. One obtained an abnormal yellow-blue diagnosis.

Eighty-six percent of CVD candidates passed the initial operational ATCOV, substantially higher than we expected from the sample of CVD subjects from Study One. However, ATCS candidates may self-select (or alternatively, somewhat screened by other testing or training processes) for color vision relative to the population of individuals with CVD. In that sense, they may differ from the CVD subjects of Study One in one important aspect. Individuals with greater CVD may choose not to pursue or may be discouraged through training and

testing for occupations known to make substantial use of color coding. If so, candidates would tend toward lower color threshold means and reduced variance and range on precision clinical tests than would CVD subjects recruited from the general population. This question can be addressed by comparing data from candidates who tested during implementation of the first operational ATCOV to subjects participating in Study Two (below) on the CAD.

All ATCS candidates who failed the ATCOV and took the CAD (n=3) had red-green threshold values greater than 12 SNU, but five of 37 who passed the ATCOV had threshold values greater than 12 SNU. Despite its precision, CAD could not fully account for scores on the initial operational ATCOV among ATCS candidates who had failed clinical screening.

Independent *Uniform Guidelines* Evaluation of the Initial Operational ATCOV

Jeffrey Hovis, an optometrist and vision scientist from the University of Waterloo, was awarded a contract to conduct an independent evaluation of compliance of the initial operational ATCOV with the *Uniform Guidelines on Employee Selection Procedures*. Dr. Hovis reviewed the test, all task analyses, and the results of Study One and produced two reports. Hovis & Ramaswamy (2010a) provided a detailed analysis and linkage of all color displays and their associated air traffic tasks. Hovis & Ramaswamy (2010b) examined *Uniform Guidelines* compliance in light of those analyses. In general, they found high content validity for the alert detection and weather identification subtests, but identified improvements required for the radar identification subtest. In addition, they recommended deletion of the radar multitasking subtest unless it could be shown to provide statistical improvement over use of the radar identification subtest alone, and deletion of the URET subtest because all uses of color were redundantly coded with position on the display. That level of redundancy facilitates training and performance and is inappropriate for selection on the basis of color vision ability. Hovis and Ramaswamy made the following recommendations to improve the content validity of the ATCOV, and we responded to each:

1. *Document the RGB color settings used in ATCOV, ARTS, STARS, and DSR and verify the chromaticity coordinates and luminances used on the test monitor are within the range of chromaticity coordinates and luminances measured on their respective displays.*

We concurred with this assessment as an issue of documentation. We accomplished this work for the initial operational ATCOV by using field-display-specified RGB values and measuring and verifying resulting chromaticities on the testing displays. For the second operational ATCOV documented below, we measured chromaticities of colors used on field displays, assembled a table of field chromaticities, and manipulated RGB values to produce the same chromaticity values on the calibrated Cathode Ray Tube (CRT) display used for CAD testing. This allowed us to move from testing using Liquid Crystal Display (LCD) to CRT displays and ensure ongoing display calibration. The ATCOV RGB values that produce field display chromaticities on the CAD CRT are shown in Table 6.

2. *Document that font size and style used in the test are representative of those used in the field.*

Table 6. ATCOV Chromaticity Specifications

Function	Optimized CAD RGB			Sample	Field Chromaticities		
	R	G	B		Y	x	y
Background	1	14	7		0.25	0.4015	0.3619
Owned DB	170	170	170		53.80	0.2769	0.3025
Alert DB	140	0	0		9.19	0.6222	0.3376
Pointout ID	156	152	0		40.68	0.3934	0.5178
Unowned DB	47	140	93		29.67	0.2782	0.4205
Highlight DB	0	150	146		34.17	0.1979	0.2526
STARS WX1	42	63	104		7.95	0.2149	0.2057
STARS WX2	80	73	55		11.35	0.3590	0.3796
ACD WX1	62	73	88		8.85	0.2689	0.2849
ACD WX2	107	70	9		12.03	0.4873	0.4402
ACD WX3	108	18	11		5.54	0.6039	0.3356
DSR WX1	4	6	92		1.60	0.1550	0.0700
DSR WX Cyn	0	84	75		10.40	0.2170	0.3320

We concurred with this assessment as a documentation issue. The fonts were specified by Kenneth Allendoerfer (William J. Hughes Technical Center) as the closest Windows approximation of deployed fonts (Lucinda Console 8pt.).

3. *Revise the radar identification practice and testing screens such that each test screen includes only the STARS or ARTS display data block and their respective colors. Although there is substantial overlap in the color codes used for each display, there are some differences in the each display's color set. There are also differences in how identical colors are used in each display in terms of the redundancies. Although the differences may appear minor to a person with normal color vision, performance of an individual with a color vision defect on a test which intermixes the two formats could be different than what would occur in the actual work environment on a single display.*

We incorporated this recommendation into the specifications for the second operational ATCOV. Each practice and test screen was specified to include target and distracter datablocks from only one system (ARTS, STARS). This increased the total number of screens on the Radar Identification subtest from 8 to 16.

4. *Ensure that the flashing data block distracters in the alert location subtest are consistent with the ARTS and STARS displays. For example, the yellow data blocks cannot flash when simulating an alert on the ARTS display; however, the green and white data blocks can flash.*

We incorporated this recommendation into the specifications for the second operational ATCOV. All flashing text, whether on alerted or distracter datablocks, conforms to their use on ARTS and STARS displays.

5. *Add a description of the redundant cues that are present to each instruction page of the Radar Identification and Alert Detection subtests. The non-color clues may, or may not be sufficient to help in identifying the data block color.*

We incorporated this recommendation into the specifications for the second operational ATCOV. Each instruction screen points out redundant coding that might assist the candidate in selecting or identifying the correct datablocks.

6. *Confirm the DSR moderate intensity weather color. There is a discrepancy between the different references, the ATCOV, and the actual DSR displays as to whether the color is blue or purple. The discrepancy may be a result of the system self-adjusting the color palette when a brighter background is used.*

The RGB values in the initial operational ATCOV were based upon chromaticities for this color measured at the Air Traffic Academy in Oklahoma City on DSR displays. Dr. Hovis' explanation of the discrepancy relative to his field observations is likely correct. In September 2009, Alan Poston (contractor to the Human Factors Research and Engineering office) obtained from the system manufacturer and provided to us chromaticity specifications for DSR. We specified for the second operational version RGB values producing these chromaticity values on the calibrated CRT monitor.

7. *Remove the URET color identification subtest due to the extensive redundancies provided for critical color codes. Interpreting the conflict warnings can be easily done based on position and brightness differences.*

We immediately revised and deployed the initial operational ATCOV software to incorporate this recommendation. The scoring software was revised to ignore data from this subtest, and all regional flight surgeons were instructed to cease testing on this subtest. We verified that no ATCS candidates had been disqualified for failing solely this subtest. While completing these analyses, we also determined that no unique failures resulted from the *Radar Multitasking* subtest in either the validation sample or candidates tested since implementation. As a result, we revised the scoring software to ignore data from this subtest and instructed all regional flight surgeons to also cease testing on this subtest. The second operational ATCOV was specified to be composed of only three subtests: Radar Identification, Alert Detection, and Weather Identification.

8. *Include naive subjects with normal color vision in the validation sample for the final version of ATCOV to ensure the cut-scores are appropriate to naive candidates who have not had any ATC training.*

We collected data for the second operational ATCOV from both NCV and CVD subjects who were naive relative to air traffic procedures and displays.

Additional Problems Requiring Correction

During implementation of the initial operational ATCOV, the regional flight surgeons identified some undesirable behaviors among ATCS candidates that were not observed during Study One, and they requested software features to overcome them. Candidates appeared to over-practice, perhaps to the point of fatigue, without any evidence of benefit. That is, those candidates who practiced the most tended to score most poorly on the subtests. At the regional flight surgeons' request, we specified software limitations of two practice sessions prior to each first subtest attempt and one additional practice session prior to retesting after a failed attempt.

Candidates also appeared to make use of the instruction-review option on the radar and weather identification subtests to gain additional time to complete test screens, again without evidence of benefit. Candidates who attempted this appeared to score most poorly on the subtests. If it were successful, such behavior would tend to pass candidates who would not be successful and would present a safety risk when actually controlling air traffic. We implemented software changes that eliminated the option to return to the instruction screen once the candidate made one selection on a test screen and to re-randomize data or weatherblock presentation on the test screen upon return from the instruction screen, if the return option was employed. This modification removed any possible advantage that might be gained by returning to the instruction screen multiple times, other than the intended allowance that a candidate might forget which data or weatherblock type he or she was instructed to search for on the test screen.

We and the regional flight surgeons were concerned for test security. The initial operational ATCOV was a fixed-format test. That is, each test screen was presented to each candidate in the same order, and correct and distracter datablocks on each individual screen always appeared in the same positions. In theory, an answer key could be developed if multiple candidates correctly memorized and published a key. This was unlikely in practice because of the sheer number of test responses required and small number of candidates who would be tested. However, if a test copy were compromised, creating a memorizable key or practicing the test to mastery would not be outside the realm of possibility. The regional flight surgeons requested and we specified software to fully randomize all test screens on the second operational ATCOV. Each test screen on the Radar and Weather Identification subtests is randomly generated when the candidate completes review of the instruction screen. The Alert Detection subtest consists of 10 fixed practice and 100 fixed scored screens, but order of presentation is determined by random selection from unpresented screens each time the candidate responds to a test screen.

The regional flight surgeons requested integration of the testing and scoring software into a single package that would manage a candidate through practice, testing, and re-testing if required, with minimal oversight by a proctor. We concurred and integrated all functions in the second operational ATCOV.

Specification of the Second Operational ATCOV

The second operational ATCOV was composed of three scored subtests: Radar Identification, Alert Detection, and Weather Identification. Testing and scoring were accomplished from a base screen from which instructions, limited practice opportunities, two test attempts, and scoring and reporting output were completed. Each subtest was specified to implement all recommendations made by Hovis & Ramaswamy (2010b) to conform to requirements of the *Uniform Guidelines*.

Study Two – Validation of Second Operational ATCOV

While revisions made to respond to the *Uniform Guidelines* evaluation and to incorporate features requested by the RFSs did not call into question the basic validity of the initial operational ATCOV, the revisions required testing of additional subjects with NCV and CVD to ensure that subtest reliability was retained and that cut-scores remained appropriate for the second operational version. Study Two was conducted to assess the reliability of the subtests, establish performance norms for NCV subjects on each revised subtest, determine cut scores to be applied in occupational testing, and examine the effect of testing upon a sample of CVD subjects on the second operational ATCOV. Cut scores were revised to ensure that a candidate with a CVD could discriminate critical display information communicated using color and redundant coding as well as NCV candidates.

In addition, we used several analyses to gauge the dimensions underlying performance on subtests to assess the degree to which constructs of color vision ability (normal/deficient or red-green/yellow-blue) determine subtest performance. We also collected several clinical measures of color vision ability among all subjects to bolster evidence of construct validity for ATCOV subtests.

Method

Data collection was accomplished among 102 volunteer subjects recruited by a subject contractor from the general population of the Oklahoma City, Oklahoma area. Subjects were naive with respect to characteristics of air traffic control and were compensated for their participation. The contractor recruited 50 subjects self-identified as having normal color vision and 52 self-identified as having a color vision deficiency.

Testing made use of a multi-purpose testing laboratory at CAMI. ATCOV testing was accomplished using the CAD monitors; ATCOV RGB settings were adjusted to produce on the CAD display chromaticity values measured on field air traffic displays as described above. The laboratory was equipped with overhead tungsten incandescent office lights, which were illuminated during testing, producing 110 cd/m^2 at the display with a chromaticity equivalent to standard light source A.

Prior approval for all procedures and use of human subjects was obtained from the FAA Institutional Review

Board. Informed consent was obtained prior to participation and subjects were free to withdraw from the project without consequence at any time.

All subjects were between the ages of 18 and 30 and were screened using a Bausch and Lomb Orthorater for visual acuity of 20/30 or better in both eyes with corrective lenses, if required. All subjects completed the second operational ATCOV twice. Practice opportunities were provided and limited by software as previously described. Additionally, CAMI personnel gave subjects multiple color vision tests, using the Nagel Anomaloscope, an experimental cockpit colors test, the Dvorine PIP test, initial operational ATCOV, Aviation Lights Test (ALT), Colour Assessment and Diagnosis (CAD), Cone Contrast Test (CCT), and Signal Light Gun Tests. Additional testing of selected individuals included identification of colored lights (incandescent and light emitting diodes; LED) and several commercially available color vision screening tests, such as the OPTEC 5000, Ishihara-38 PIP, OPTEC 900, Waggoner PIP, and the Titmus i400.

For analysis purposes, color vision was examined using the Dvorine, CCT, and CAD tests. Subjects were classified as NCV or CVD using the CAD test, regardless of their self-identified CV status. This method differed from Study One, which classified CVD subjects by the Nagel anomaloscope and accepted that ATCS trainees had been screened for normal color vision. We selected CAD classification because:

- it tests for both red-green and yellow-blue deficiencies
- diagnosis does not require an expert test administrator
- it provided comparability to CVD ATCS candidates tested during implementation of the initial operational ATCOV.

Rodriguez-Carmona (2006) found strong agreement between CAD and the Nagel Anomaloscope for red-green deficiency ($\text{Kappa}_{n=224}$ = .99).

Subjects with red-green threshold scores on the CAD of 1.7 or less and yellow-blue threshold scores of 1.8 or less (the values at which CAD diagnoses "potential" deficiencies) were classified as having normal color vision. By these criteria, 59 subjects had NCV and 42 were CVD. One subject was unclassifiable by CAD and was excluded from all subsequent analyses. Kappa values between CAD classification at these cutoffs and Dvorine and CCT classifications as NCV or CVD were .63 and .75, respectively. CAD red-green (RG) threshold scores were correlated .73 with number of correct Dvorine responses and .64 and .68 with CCT correct Red and Green responses. CAD yellow-blue (YB) threshold scores were not significantly correlated with correct Dvorine responses but correlated .50 with CCT correct Blue responses. CAD classification differs from a pass-fail classification on the Dvorine by assessing YB deficiency. Initial analyses indicated that eight of the 69 subjects who passed the Dvorine had yellow-blue CVD, which is not tested by the Dvorine. To obtain a normal color vision sample, we believed screening of yellow-blue deficiencies was necessary. The sample included 23 deutans, 11 protans, 1 tritan, and 7 subjects evidencing both red-green and yellow-blue deficiencies.

Results

Reliability analysis. It was more difficult to calculate internal consistency on the Radar and Weather Identification when subtest screens were randomized. In the initial version, each correct answer could be treated as an item. In the second version, each *screen* had to be treated as an item (because screens varied across subjects in numbers of correct and distracter data or weatherblocks and in location within each screen). This substantially reduced the number of items, from 10 items per target type and a minimum of 40 items per subtest to 16 screens per subtest. Reliability calculations are greatly influenced by the number of items. More significantly, on six Radar Identification and two Weather Identification screens, NCV subjects made no errors; this results in zero variance on these items, making alpha incalculable among NCV subjects for a scale including those items. However, a lack of variance because of correct responses by NCV subjects is a problem solely for internal consistency *calculation* – from a testing perspective, it is evidence of a ceiling effect among NCVs, who *should* have no difficulty with the items.

Alert Detection was not affected by item-reduction due to randomization because each of the 100 screens had always included a single correct item. However, all NCV subjects correctly answered 59 items, making alpha incalculable for the 100-item subtest.

Test-retest values and alphas were calculable for all subtests when both NCV and CVD subjects were included in a single analysis. The resulting internal consistency values obtained are shown in Table 7.

All calculable values were in the acceptable range and were comparable to those observed in Study One. Kappa for pass/fail on first and second attempts was .64.

Table 7. Internal Consistency Values for ATCOV Subtests

Subtest	*a* – all subjects	*a* – CVD	*Test-retest – all subjects*
Radar Identification	.88	.88	.95
Alert Detection	.97	.98	.80
Wx. Identification	.83	.88	.89

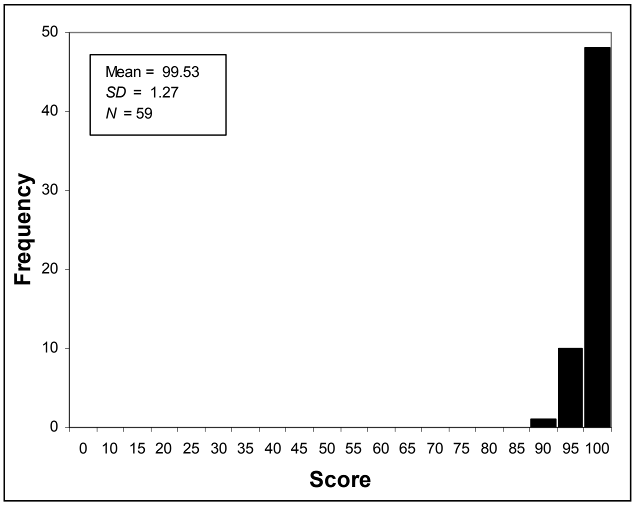

Figure 6a. Radar Identification Scores Among NCV Subjects

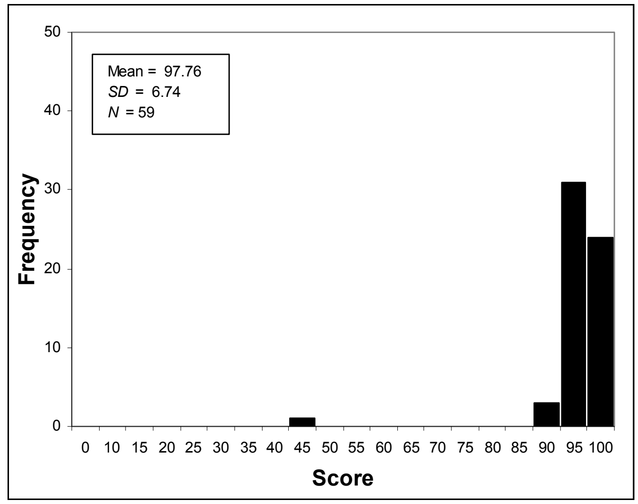

Figure 6b. Alert Detection Scores Among NCV Subjects

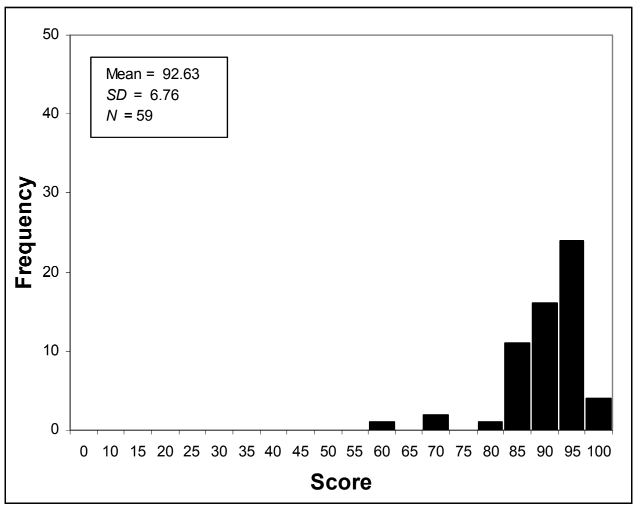

Figure 6c. Weather Identification Scores Among NCV Subjects

Normal color vision subjects. Air traffic naive NCV subjects who participated in Study Two were comparable to ATC candidates who participated in Study One when compared along the three surviving subtests of the initial operational ATCOV. Mean scores differed significantly only for Alert Detection scores, with Study Two participants averaging .26 standard deviations higher. Scores for Radar and Weather Identification did not differ between the two samples. Overall, 95% of NCV Study Two subjects passed the initial version, which surpasses results from Study One, probably because the CAD was used to classify individuals as NCV or CVD versus classification by medical clearance for the NCV group in Study One.

Distributions on the second operational ATCOV met expectations for each subtest, with scores concentrated at the upper range and tailing off sharply towards lower scores, as shown if Figures 6a through 6c.

Mean, median, standard deviation, and fifth percentile score for each subtest is documented in Table 8.

Distributions also mirrored those of the research and initial operational versions of ATCOV, except that scores on the second operational ATCOV were higher for Radar Identification.

Cut-scores were reset relative to the initial operational ATCOV only where data from these 59 NCV subjects differed significantly from the larger previous sample (e.g., note that the fifth percentile is the third lowest-scoring subject in a sample of 59). Thus, the cut-score for Radar Identification was increased to 95 because the NCV sample mean for the second operational ATCOV differed by more than half a standard deviation ($t = 5.86$; $p < .01$) from the initial operational version. Passing score for the Alert Detection subtest remains at 89. Passing

Table 8. Descriptive Statistics for ATCOV Subtests Among NCV Subjects

Subtest	Mean	Median	Std. Dev.	Fifth Percentile
Radar Identification	99.53	100.00	1.27	95
Alert Detection	97.76	99.00	6.74	89
Wx. Identification	92.63	93.75	6.78	80

score for the Weather Identification subtests remains at 80. No cut-scores were set for reaction time. Applying the revised cut scores to the NCV sample, 91.5% passed all subtests on their first attempt and 96.6% passed all subtests after two attempts. Two NCV subjects (3.4%) failed at least one subtest on both attempts; both failed Alert Detection. These subjects were difficult to interpret; they are NCV (passing Dvorine, CAD, CCT, and the initial operational ATCOV) but could not pass the second operational ATCOV. To the good, ATCOV failures among NCV subjects may be expected simply because we set a cut score at the fifth percentile on the first testing attempt. These subjects were below the fifth percentile on both attempts. In that sense, they may be a simple statistical artifact. To the bad, we would expect all NCV ATCS *candidates* to be able to pass the test after two attempts and observed this among ATCS trainees who took the initial operational ATCOV twice in Study One. Importantly though, these *subjects* were not *candidates*; they were recruited from the Oklahoma City area without requirement of academic credentials or other screening normally required of candidates. Candidates are screened for a variety of abilities for which our subjects were not. In practice, these two subjects would never be tested on the ATCOV if they presented as candidates. Having passed all normal CV screens, they would be medically cleared without occupational testing.

Color vision deficient subjects. CVD subjects who participated in Study Two were compared with those who participated in Study One on the three surviving subtests of the initial operational ATCOV. Mean scores differed significantly for Radar Identification and Alert Detection, but Study Two participants averaged .61 standard deviations *higher* on Radar Identification and 1.25 standard deviations *lower* on Alert Detection. Scores for Weather Identification did not differ between the two samples. Overall, 44.2% of CVD Study Two subjects passed the initial version, which surpasses passing rates observed in Study One.

In addition, CVD subjects recruited for Study Two were compared with ATCS candidates who had completed CAD testing. Study Two subjects had greater variance (standard deviations of 6.79 versus 3.74 for red-green threshold and 1.58 versus .34 for yellow-blue threshold; Fs = 3.30 and 21.88, respectively, $p<.01$) and a more extreme range of scores (38% of Study Two CVD subjects had greater red-green thresholds than the most extreme CAD-tested ATCS candidate and 14% had greater yellow-blue thresholds) than candidates who had completed CAD testing. Study Two subjects had greater mean yellow-blue threshold scores (1.81 versus 1.05, $t=2.49$, $p<.01$), but mean red-green thresholds were not significantly different. Taken together, this suggests that candidates may be somewhat self-selected for color vision (or alternatively, that other selection tests and training programs had exercised some degree of color vision screening), such that persons with more extreme deficiencies are less likely to apply, complete training, or be selected pending medical fitness evaluation.

Among CVD subjects, distributions for each subtest of the second operational ATCOV were consistent with expectations from the research and initial operational versions of ATCOV: on average, CVD subjects do not score as well on any subtest, but a significant proportion can discriminate critical datablocks and weatherblocks as effectively as NCV subjects, as shown in Figures 7a to 7c. Comparison to Figures 6a to 6c reveals the distributions to be shifted to the left among CVD subjects. There were fewer perfect scores and substantial numbers of scores in the 60% to 80% range.

Mean, median, standard deviation, and passing rates on the first and second attempt for each subtest are documented in Table 9.

Overall, 46.5% of CVD subjects passed all subtests on their first attempt; 58.1% passed after two opportunities on all subtests. Alert Detection was the most difficult subtest. This is expected given the criticality of alerts and justifiable limitations on presentation time. This passing rate is equivalent to that of the initial operational ATCOV among these subjects (44.2% on the sole attempt), higher than that observed among CVD subjects in Study One (32%), but less than that among field candidates who had failed clinical testing (86%). Passing rates differed by type of diagnosis: 55% of 11 protans passed after two attempts, 70% of 23 duetans passed, the sole tritan failed, and 43% of those with both RG and YB deficiencies passed.

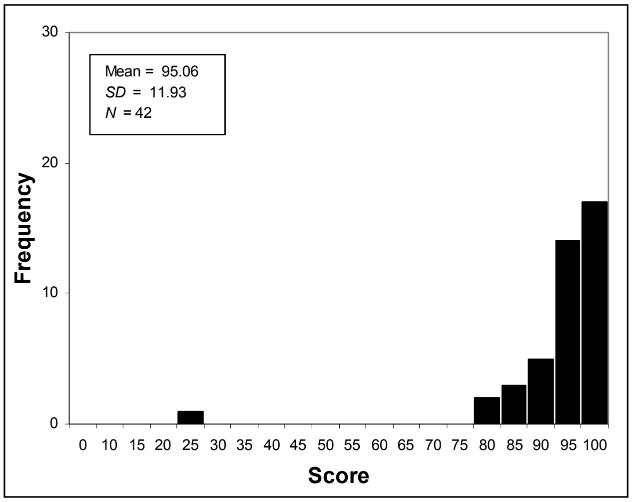

Figure 7a. Radar Identification Scores Among CVD Subjects

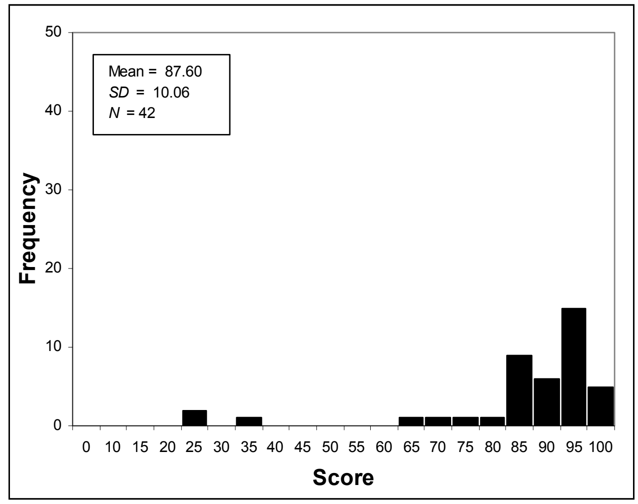

Figure 7b. Alert Detection Scores Among CVD Subjects

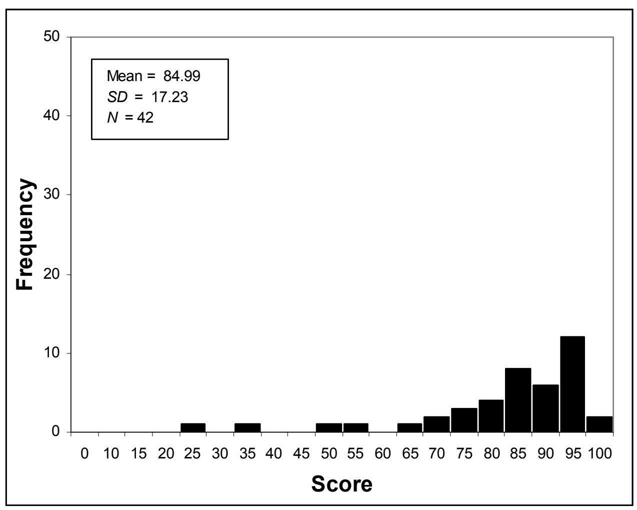

Figure 7c. Weather Identification Scores Among CVD Subjects

Table 9. Descriptive Statistics for ATCOV Subtests Among CVD Subjects

Subtest	Mean	Median	Std. Dev.	Pass 1st Attempt	Pass 2nd Attempt
Radar Identification	95.06	98.75	11.93	72%	79%
Alert Detection	87.60	93.50	18.06	60%	70%
Wx. Identification	85.00	89.84	17.23	74%	83%

Analyses contrasting subtest performance of NCV and CVD subjects. Compared with NCV subjects, CVD subjects in Study Two scored significantly lower on all subtests on their first attempt, as shown in Table 10.

On their second attempt, CVD subjects made up significant ground on the Radar Identification subtest, due in part to slightly lower scores and greater variability among NCVs, thus reducing the effect size between the groups, as shown in Table 11.

Effect sizes (*d*) were smaller than those found in Study One. A Kappa of .46 was obtained between CV classification by the CAD and ATCOV pass-fail results on the first attempt and decreased to .41 after two attempts. The former value is somewhat lower than found using classification by Nagel Anomaloscope in Study One. Kappa values of .38 and .33 were obtained for ATCOV with Dvorine and CCT CV status, respectively. Correlations between ATCOV subtests and CAD, Dvorine, and CCT scale scores were low except for Alert Detection with CAD thresholds, as shown in Table 12.

Test improvements made to fully comply with the *Uniform Guidelines*, reduce potential confusions of display formats, and base norms on air traffic naive subjects resulted in the same patterns of passing and failing between NCV and CVD subjects, but appeared to reduce the sensitivity of the test to CVDs.

Analyses examining construct validity. We replicated the three analyses from Study One by combining the NCV and CVD samples to assess construct validity. Importantly, these analyses relied upon three, rather than five, subtests and included less than half of the sample size of Study One.

1. We contrasted NCV versus CVD subjects using discriminant analysis of second operational ATCOV subtest scores. The discriminant function gave greatest weight (.74) to Alert Detection, .43 to Weather Identification, and .22 to Radar Identification ($p<.01$) and correctly classified 76% of subjects by NCV versus CVD status. If subjects were scored by this function, 91% of NCV subjects and 45% of CVD subjects would be classified in a NCV group, and 9% of NCV and 55% of CVD subjects would be classified in a CVD group. This is equivalent to passing rates observed for the test and resulted in a Kappa of .49 for CV status and .86 for passing or failing the initial operational ATCOV. With the caveat of fewer predictors and a smaller sample size, this analysis reinforces the conclusion that the second operational ATCOV appears to

Table 10. Comparison of Subtest Scores by NCV and CVD Subjects on First Attempt

Subtest	NCV	CVD	d	p
Radar Identification	99.53	95.06	.56	<.01
Alert Detection	97.76	87.60	.75	<.01
Wx. Identification	92.63	85.00	.60	<.01

Table 11. Comparison of Subtest Scores by NCV and CVD Subjects on Second Attempt

Subtest	NCV	CVD	d	p
Radar Identification	98.52	95.29	.33	<.10
Alert Detection	98.78	90.93	.79	<.01
Wx. Identification	94.74	88.15	.55	<.01

Table 12. Correlations Among Color Vision Test Scores

ATCOV	CAD RG	CAD YB	Dvorine	CCT Red	CCT Green	CCT Blue
Radar Identification	-.17	-.19	.14	.14	.15	-.035
Alert Detection	-.39*	-.54*	.08	.20*	.20*	.02
Wx. Identification	-.10	-.16	.11	.07	.17	.10

*p < .05

have become somewhat less sensitive to color vision differences when its content validity was improved.

2. We cluster-analyzed the subjects, ignoring NCV versus CVD group membership. This approach was not successful in this sample. A two-cluster solution separated one subject from all others. Adding clusters slowly separated individuals from the larger cluster. We attempted a further analysis using the cluster centroids from Study One on the three surviving subtests as initial centers for classifying Study Two subjects. This analysis showed little improvement, grouping all but one of the NCV subjects with 78% of CVD subjects and 22% of CVD subjects with a single NCV subject. This resulted in a Kappa of .22 for CV status and .46 for passing or failing the initial operational ATCOV. Cross-tabulation of predicted CV status by the discriminant analysis with cluster membership resulted in a Kappa of .44.
3. We applied factor analysis (Principal Components extraction with Varimax rotation) to the three subtests to assess the dimensionality of constructs underlying performance on the second operational ATCOV. Analysis of Eigenvalues and alternative numbers of extracted factors suggested that two factors accounted for response variance. A single-factor solution accounted for 59% of variance and weighted Radar Identification .87, Alert Detection .49, and Weather Identification .88. Given that the greater-weighted subtests were less tied to color vision, this approach suggests an unidentified second factor. A forced twp-factor solution accounted for additional variance (88%), grouping Radar and Weather Identification on one factor and Alert Detection (most closely tied to color vision) on another. This solution seemed to focus on testing methodology differences, separating identification from time-critical alert detection functions.

Seeking clarification of factor structure, we added CAD red-green and yellow-blue threshold scores to a factor analysis of ATCOV subtests. A two-factor solution accounted for 68% of variance and weighted Alert Detection with both CAD thresholds on one factor, and Radar and Alert Detection on the other. This solution highlights the impact of color vision on Alert Detection noted in other analyses. A three-factor solution accounting for 85% of variance seemed to offer the greatest clarity, suggesting red-green and yellow blue dimensions, with Alert Detection weighted moderately (.40) with red-green and strongly (.77) with yellow-blue threshold scores. Radar and Weather Identification were loaded on the third factor. This solution seems the best fit with what is known about color vision, but re-emphasizes the reduced impact of color vision on test scores and implies that clinical screening for yellow-blue deficiency may be necessary, since that relatively rare type of deficiency appears to be at least moderately correlated with a reduced ability to detect alerts.

One final factor analysis was conducted to further explore the apparent reduction of impact of color vision ability on the second operational ATCOV. Scores from the three surviving subtests of the initial operational ATCOV and CAD red-green and yellow-blue threshold scores were factor analyzed. A two-factor solution appeared to be the best fit, accounting for 72% of the variance. The rotated solution weighted Radar Identification .92, Weather Identification .73 and CAD RG threshold -.69 on one factor, and Alert Detection -.88 and CAD YB threshold .92. A forced three-factor solution accounted for 88% of variance and weighted CAD YB threshold (.94) with Alert Detection (-.86) on one factor, CAD RG threshold (-.93) with Radar Identification (.69) and Alert Detection (.30) on a second factor, and Weather Identification (.96) and Radar Identification (.62) on a third factor. The first and second operational versions behaved differently with respect to color vision. Alert Detection was strongly associated with YB and moderately associated with RG threshold scores on both versions of ATCOV, but both Identification subtests were much less sensitive to color vision ability on the second operational ATCOV.

Two additional discriminant analyses were conducted, making use of CAD and CCT data to predict passing or failing the second operational ATCOV after two attempts. The CAD function correctly classified 79% of cases, while the CCT function correctly classified 75%. However for both functions, discrimination of CVD subjects who passed from those who failed was little better than chance. Neither CAD nor CCT scores could account for ATCOV outcomes, presumably because of the way redundant coding provided in the operational environment was represented in the occupational test.

Discussion

The Study Two results provided evidence that the second operational ATCOV subtests were reliable, established performance norms for NCV subjects on each subtest, determined cut scores to be applied in occupational testing, and examined the impact of testing upon a sample of CVD subjects. In general, the subtests were internally consistent when both NCV and CVD subjects were included in the analysis, though internal consistency could not be assessed when NCV subjects

were analyzed separately due to ceiling effects. Scores were stable among both NCV and CVD subjects. Cut scores were set to ensure that a CVD candidate who passed the test could discriminate critical information communicated using color as well as NCV candidates. Subtest scores generally separated NCV from CVD subjects, but identified fairly substantial numbers of CVD subjects who could discriminate critical information communicated using color. Having incorporated changes identified in the independent assessment of *Uniform Guidelines* compliance, we can be confident that test construction adequately sampled critical information communicated using color on critical displays. With those changes, the subtests continue to function as desired and expected.

Discriminant, cluster, and factor analyses suggested that ATCOV subtests measure color vision ability but became less sensitive as content validity was improved. Subtest scores separated NCV from CVD subjects on average, but substantial numbers of CVD subjects were able to pass ATCOV. By including precise measures of color vision ability (CAD and CCT), we learned that subtest scores appear to be tied to both red-green and yellow-blue dimensions of color vision. This provided further evidence of construct validity for the ATCOV, even though the effect of color vision deficiency on subtest scores weakened as content validity for ATCS tasks was improved.

Deployment of the Second Operational ATCOV

Deployment was initiated in August and completed in November of 2010. The process consisted of traveling to each regional flight surgeon or medical field office location, installing all CAD updates on the test computer, calibrating the CRT monitor, removing the LCD monitor, installing the ATCOV update, and training proctors in how to use the new software. CAD testing of candidates was then discontinued for lack of a further research purpose because our Study Two analyses showed that precision clinical testing did not adequately predict occupational test scores. We specified a standard lamp and tungsten fluorescent bulbs to testing locations to provide ambient illumination comparable to conditions under which the second operational ATCOV was validated. Test administrators were provided with a luminance meter and instructed to place the lamp in a location that produced 110 cd/m^2 at the display. Chromaticity was controlled by bulb specification to produce a chromaticity equivalent to standard light source A.

CONCLUSIONS AND RECOMMENDATIONS

Importance of Color Vision to Controller Task Performance

Color vision ability remains critical to the provision of air traffic services in the National Airspace System. Using cognitive analyses of ATCS tasks (Nickels et al., 1995; American Institutes for Research, 2006a, 2006b, 2010a), our preliminary analysis identified critical job activities and tasks for which color vision is critical to successful performance due to color coding used on displays to communicate necessary information. Hovis and Ramaswamy (2008a) verified and further explicated these linkages. As required under the *Uniform Guidelines*, this linkage establishes the relationship of the test to the job in terms of the likelihood that a person with CVD will be able to perform the critical or important job tasks requiring color/type discrimination. Task analysis linkage emphasized that ATCSs must be able to discriminate 1) among datablocks coded in color to indicate whether they are under the control of the candidate (owned), under control of someone else (unowned), being pointed out to the candidate (pointout), or in alert status (alert), and from datablocks coded for non-critical purposes (e.g., optional highlighting); and 2) among each level of weather severity communicated within a display type (ARTS, STARS, DSR). Additionally, ATCSs must be able to rapidly detect and accurately interpret datablocks in alert status to prevent collision of associated aircraft with another aircraft, terrain, or obstacles.

Addressing the Occupational Testing Requirement

Allowing color vision deficient controllers to qualify was mandated by the courts to comply with the provisions of the Rehabilitation Act and American with Disabilities Act, as not all color vision deficient individuals are unable to see and perform critical functions. The FAA must assess which individuals with a color vision deficiency can reliably perform critical controller tasks. The Federal Air Surgeon and Human Factors Research, Engineering, and Development office tasked the Civil Aerospace Medical Institute to develop, validate, and implement an occupational test for ATCS job candidates who are identified as having a color vision deficiency by clinical screening during the pre-employment medical examination. Xing (2008a, 2008b) completed a research prototype, and we were tasked to implement an operational version compliant with all applicable standards, principally the *Uniform Guidelines,* to serve as an occupational test for color vision deficient controller candidates. In contrast to the research ATCOV, the second operational ATCOV complies with the *Uniform Guidelines* reporting requirements for both

content and construct-oriented validity. Evidence of content validity for ATCS duties is provided through direct sampling of form and content of critical display data. Evidence of construct validity is provided by correlation with CAD and CCT threshold scores, which precisely measure color vision ability.

This resulted in a job sample test closely tied to critical tasks communicated using color on air traffic displays. ATCOV makes use of display formats and color chromaticities deployed for critical information on critical displays, as defined by published analyses of ATCS tasks. Its items are isomorphic with datablocks and weather depictions deployed on ARTS, STARS, and DSR displays in terminal and en route facilities.

Constructs Underlying ATCOV Test Performance

Discriminant and factor analyses suggested dimensions underlying the ATCOV. To have construct validity for color vision ability, scores on the ATCOV should be moderately correlated with scores on clinical tests, but recommended outcomes should differ somewhat. Persons with mild to moderate CVD identified by most clinical screening tests have been shown to be able to pass the ATCOV if they can adequately discriminate informational coding with the current use of a limited palette of colors or by using redundant coding. This was observed in correlations with CAD threshold scores and in exploratory factor analyses. Analyses indicated that severity of red-green and yellow-blue deficiency is a good, but imperfect predictor of passing ATCOV. Alert Detection, which is the most difficult subtest, is moderately correlated with CAD threshold scores. Radar and Weather Identification subtest scores were less correlated with threshold scores. This is an appropriate set of outcomes. ATCOV is administrated as an occupational test given to candidates who have failed clinical color vision screening. Though it presents data and weather blocks in the colors used in the field, ATCOV is not a test of color vision, *per se*. Instead, ATCOV tests a candidate's ability to make the color discriminations required to perform the job.

As the test evolved from the research to the initial and second operational versions, the results revealed a tradeoff of construct for content validity. As field display content was increasingly closely represented, more individuals with CVD were able to pass the test. This is appropriate for an occupational test; those who can discriminate required information despite their color vision deficiency should be medically qualified for the position.

Clinical Screening for Yellow-Blue Color Vision Deficiency

Further consideration is required for yellow-blue CVD. Study Two revealed that CAD yellow-blue threshold scores were moderately correlated with Alert Detection. Currently, ATCS candidates with yellow-blue deficiencies can pass clinical screening, if given a test that does not include yellow-blue screening plates, such as the Dvorine or Ishihara PIP tests. Some of these candidates will likely have difficulty with alert detection in training and on the job. Under current procedures, these candidates would be medically qualified without occupational testing.

What safety risk might this represent? To the good, yellow-blue is the rarest form of CVD. Tritans typically account for less than 1% of the population, while protan and deutan deficiencies affect about 8% of males. However, yellow-blue deficiency can be acquired due to a variety of medical conditions or may appear with the use of a number of medications, and most individuals show some degree of deficiency with age, due to corneal yellowing (Yates, Diamantopoulos, & Daumann, 2001). Aging is highly unlikely to be the cause of deficiencies observed in Study Two or in the ATCS *candidate* population because all subjects were between 18 and 30 years old and statutory requirements limit the hiring age for controllers to not exceed 30 years. One possible solution would be to require ATCS screening using tests that assess yellow-blue deficiency (such as the Waggoner HRR or the Richmond HRR 4th Edition), followed by occupational testing using ATCOV among those who fail clinical screening.

Necessity of Display Standards for Color Use

Future challenges will surround the stability of color use on new systems and displays. For example, the Ocean 21 system was deployed to en route centers that provide oceanic air traffic services while ATCOV was under development. It adds at least one critical color and uses different presentation formats with additional redundant coding. Additionally, the En Route Automation Monitoring (ERAM) system is being tested in two en route facilities. When fully deployed, it will replace DSR. ERAM uses a large number of colors, but a task analysis linkage is not yet available to determine which are used for critical information. Both of these systems will require updating the ATCOV and may expose cleared candidates to future inability to perform safely (c.f., Crutchfield & Lowe, 2010).

Additional new ATC systems will present the same risks if a standard color palette is not established. Risks to safety could increase if the critical color vision palette were allowed to change in a manner resulting in cleared ATCSs becoming unable to discriminate critical information

communicated using different colors, if undetected. Increased risk could also occur by disqualifying highly trained ATCSs who could no longer accomplish a critical task or activity due to CVD, thereby reducing the overall experience level of the workgroup. Candidates cleared by ATCOV have demonstrated the ability to discriminate the currently deployed color palette with use of redundant coding of some colors.

This cohort of cleared color vision deficient controllers will become a *de facto* constraint on future displays, but we should act to formalize those constraints. We should ensure that future displays use colors that are readily discriminable by medically cleared candidates. Future displays need to be restricted either to the current color palette (which is untenable, given known deployments forthcoming) or to a standard, limited color set, defined in chromaticity, that cleared candidates can be shown to discriminate.

RECOMMENDATIONS

- ATCOV should serve as the occupational test for future candidates who fail clinical color vision screening. ATCOV supports color vision requirements presently implemented in current agency orders (FAA, 1996). An independent evaluation documents its compliance with the *Uniform Guidelines*.
- A new subtest, or modifications to existing subtests, must be developed to represent the Ocean21 system. When Air Traffic finalizes a decision to deploy ERAM to all en route facilities, an additional subtest or modifications to existing subtests must be added to the ATCOV to test discrimination of the expanded critical color palette.
- The FAA should conduct further research to develop and implement a standard color palette for future air traffic displays.

REFERENCES

American Institutes for Research. (2006a). Air traffic control job analysis: A summary of job analytic information for air traffic En Route controllers. Contractor Report. Washington, DC: Federal Aviation Administration.

American Institutes for Research. (2006b). Air traffic control job analysis: A summary of job analytic information for air traffic tower controllers. Contractor Report. Washington, DC: Federal Aviation Administration.

American Institutes for Research. (2006c). Air traffic control job analysis: A summary of job analytic information for air traffic TRACON controllers. Contractor Report. Washington, DC: Federal Aviation Administration.

Allendoerfer, K., Friedman-Berg, F., & Pai, S. (2007). Human factors analysis of safety alerts in air traffic control (Report No. DOT/FAA/TC-07/22). Washington, DC: Federal Aviation Administration.

Barbur, J., Rodriguez-Carmona, M., Evans, S., & Milburn, N. (2009). Minimum color vision requirements for professional flight crew, part III: Recommendations for new color vision standards (Report No. DOT/FAA/AM-09/11). Washington, DC: Federal Aviation Administration.

Broach, D. (2009). Assessment of documentation on the development and validation of the air traffic color vision (ATCOV) practical test relative to the reporting requirements of the Uniform Guidelines on Employee Selection Procedures. Unpublished memorandum. Oklahoma City, OK: Federal Aviation Administration.

Cardosi, K.M. & Boole, P.W. (1991). Analysis of pilot response time to time-critical air traffic control calls (Report No. DOT/FAA/RD-91120). Washington, DC: Federal Aviation Administration.

Crutchfield, J.M. & Lowe, S.E. (2010). Exploring the relationship between operational errors and color vision deficiency. Presented at The 118th Annual Convention of the American Psychological Association , San Diego, CA.

Cummings, M.L., Tsonis, C., & Xing, J. (2007). Investigating the use of color in timeline displays, (Report No. DOT/FAA/AM-07/24). Washington, DC: Federal Aviation Administration.

Equal Opportunity Employment Commission (1978). Uniform Guidelines on Employee Selection Procedures. 29 CFR § 1607; Code of Federal Regulations. Washington, DC.

Federal Aviation Administration. (1996). Air Traffic Control Specialist Health Plan. Office of Aerospace Medicine Order 3930.3a. Washington, DC: United States Department of Transportation.

Frieman-Berg, F., Allendoerfer, K., & Pai, S. (2008). Moving toward an air traffic control display standard: Creating a standardized color palette for terminal situation displays (Report No. DOT/FAA/TC-08/15). Washington, DC: Federal Aviation Administration.

Hovis, J. & Ramaswamy, S. (2010a). Relationships of the operational air traffic controllers color vision test to air traffic controllers' tasks: Review of color related tasks. Contractor Report. Oklahoma City, OK: Federal Aviation Administration.

Hovis, J. & Ramaswamy, S. (2010b). Validity of the air traffic controllers color vision test. Contractor Report. Oklahoma City, OK: Federal Aviation Administration.

International Commission on Illumination. (1931). *Commission internationale de l'Eclairage proceedings.* Cambridge: Cambridge University Press.

Mertens, H.W. (1990). Evaluation of functional color vision requirements and current color vision screening test for air traffic control specialists (Report No. DOT/FAA/AM-90-9). Washington, DC: Federal Aviation Administration.

Mertens, H.W., Milburn, N. J., & Collins, W.E. (1995). Practical color vision tests for air traffic control candidates: En route and terminal facilities (Report No. DOT/FAA/AM-95-13). Washington, DC: Federal Aviation Administration.

Milburn, N.J. (2004). A historical review of color vision standards for air traffic control specialists at automated flight service stations (Report No. DOT/FAA/AM-04-10). Washington, DC: Federal Aviation Administration.

Nickels, B.J., Bobko, P., Blair, M.D., Sands, W.A., & Tartak, E.L. (1995). Separation and control hiring assessment (SACHA) final job analysis report. Contractor Report. Washington, DC: Federal Aviation Administration.

Office of Personnel Management. (undated). Qualification standards, air traffic control series (2152). Washington, D.C. Available at: http://www.opm.gov/qualifications/standards/IORs/gs2100/2152.htm

Rabin, J., Gooch, J., & Ivan, D. (2010). Rapid quantification of color vision: The Cone Contrast Test. Investigative Ophthalmology and Visual Science. 52 (2): 816-820.

Rodriguez-Carmona, M.L. (2006) Variability of chromatic sensitivity: Fundamental studies and clinical applications. Unpublished doctoral dissertation. London, United Kingdom: City University.

Rodriguez-Carmona, M.L., Harlow, A.J., Walker, G., & Barbur, J.L. (2005). The variability of normal trichromatic vision and the establishment of the "normal" range. Proceedings of the 10th Congress of the International Colour Association. Granada.

Tanner, W. & Swets, J. (1954). A decision-making theory of visual detection. Psychological Review 61 (6): 401–409.

Xing, J. (2006a). Reexamination of color vision standards, Part III: Analysis of the effects of color vision deficiencies in using ATC displays (Report No. DOT/FAA/AM-06/11). Washington, DC: Federal Aviation Administration.

Xing, J. (2006b). Color and visual factors in ATC displays (Report No. DOT/FAA/AM-06/15). Washington, DC: Federal Aviation Administration.

Xing, J. (2006c). Color analysis in air traffic control displays, Part I. Radar displays (Report No. DOT/FAA/AM-06/22). Washington, DC: Federal Aviation Administration.

Xing, J. (2007a). Color analysis in air traffic control displays, Part II. Auxiliary displays (Report No. DOT/FAA/AM-07/05). Washington, DC: Federal Aviation Administration.

Xing, J. (2007b). Developing the Federal Aviation Administration's requirements for color use in air traffic control displays (Report No. DOT/FAA/AM-07/10). Washington, DC: Federal Aviation Administration.

Xing, J. (2008a). ATCOV, I: Development of a practical job selection test. Unpublished manuscript. Oklahoma City, OK: Federal Aviation Administration.

Xing. J. (2008b). Air Traffic Color Vision Test (ATCOV), II: Empirical verification and evaluation. Unpublished manuscript. Oklahoma City, OK: Federal Aviation Administration.

Xing, J., Broach, D., Ling, C., Manning, C., & Chidester, T. (2009). Air Traffic Color Vision Test (ATCOV), III: Validating the test for operational use. Unpublished manuscript. Oklahoma City, OK: Federal Aviation Administration.

Xing, J., & Ling, C. (Under review). Air Traffic Color Vision Test (ATCOV) II: Empirical verification and evaluation. Oklahoma City, OK: Federal Aviation Administration.

Xing, J., & Manning, C.A. (2005). Complexity and automation displays of air traffic control: Literature review and analysis (Report No. DOT/FAA/AM-05/04). Washington, DC: Federal Aviation Administration.

Xing, J. & Schroeder, D.J. (2006a). Reexamination of color vision standards, Part I. Status of color use in ATC displays and demography of color-deficient controllers (Report No. DOT/FAA/AM-06/02). Washington, DC: Federal Aviation Administration.

Xing, J. & Schroeder, D.J. (2006b). Reexamination of color vision standards, Part II. A computational method to assess the effect of color deficiencies in using ATC displays (Report No. DOT/FAA/AM-06/06). Washington, DC: Federal Aviation Administration.

Yates, J., Diamantopoulos, I., & Daumann, F. (2001). Acquired (transient and permanent) colour vision disorders. In Menu, J. et al., *Operational Colour Vision in the Modern Aviation Environment*. Neuilly-Sur-Seine Cedex, France: North Atlantic Treaty Organization Research and Technology Organization.

NOTES

1. URET testing was subsequently removed from the test as described later in the paper.
2. Initially, linkage included a focus on URET and multitasking. However, as described subsequently in the paper, after independent *Uniform Guidelines* evaluation and empirical analyses demonstrating that radar identification testing did not differ with or without multitasking, these subtests were dropped.
3. Test screen presentation time was selected based upon analysis of two critical types of alerts, low altitude and collision. Alert logic for each function is designed to give a minimum 30-second alert time (look-ahead time) to terrain/obstacle or traffic conflict, but may lose 5 seconds in a worst-case scenario due to the requirement for identification by multiple display updates before alerting. Alerts are accompanied by a 5-second aural alarm. Allendoerfer, Friedman-Berg, & Pai (2007) completed analyses of these alerts and determined that look-ahead must allow sufficient time for the controller to:
   - detect the aircraft in conflict – the subject of this subtest
   - obtain necessary information (e.g., read the altitudes from the data blocks)
   - decide what action to take
   - communicate instructions to pilots

   Following these actions, it must allow pilots within the remaining alert time to:
   - hear and acknowledge the instructions
   - implement the instructions

   And finally, time for the aircraft to:
   - respond to pilot inputs and become clear of conflict

Time requirements for most of these processes have been documented in previous research (Allendorfer et al., 2007; Cardosi & Boole, 1991). The combination of the controller detecting, obtaining necessary information, deciding on action, and communicating an instruction averages 5 seconds. Pilot acknowledgment of instructions averages 3 seconds. Lag time for pilot response to instructions averages 5 seconds. In a worst case alerting scenario (25 seconds), this would leave 12 seconds for the aircraft to respond to pilot inputs and become clear of conflict. By contrast, consider that in the American Airlines accident in Cali, Colombia (a non-radar environment) in December of 1995, the crew received a Ground Proximity Warning System terrain alert 13 seconds prior to impact. Test screen presentation time was selected to measure perception/detection time. Allowing 2 seconds for detection will ensure communication of instruction in 5 seconds when necessary to respond to a conflict alert.

4. Lighting accommodation may be appropriate for testing of on-board CVD controllers, should future color vision testing of incumbents become necessary. If so, office lighting conditions should be used for tower controllers and dim light or darkness for TRACON/ARTCC controllers.

APPENDIX A

Preliminary Linkage of Functions Using Color Coding to Critical ATCS Tasks

System	Function	Color	Redundant Coding	AIR Job Analysis Reference
DSR	Owned aircraft	white	Greater luminance	6A.01.01 (perform situation monitoring)
	Unowned aircraft	white	"R"	6A.01.01 (perform situation monitoring)
	Pointout	white	Appears when pointed out	6A.01.014.08 (issue pointouts); 6A.01.04.09 (respond to pointouts)
	Alert	white	Datablock flashes, alert code appears	6A.01.02 (resolve aircraft conflict situations)
	Weather 1	purple	None	6A.01.05 (assess weather impact)
	Weather 2	turquoise-black stippled	None	6A.01.05 (assess weather impact)
	Weather 3	turquoise	None	6A.01.05 (assess weather impact)
URET (aircraft list)	Owned aircraft	white	Greater luminance	6A.01.03 (manage air traffic sequences); 6A.01.04 (route or plan flights)
	Predicted conflict	red	Position	6A.01.02 (resolve aircraft conflict situations)
	Potential conflict	yellow	Position	6A.01.02 (resolve aircraft conflict situations)
	Wrong altitude for direction of flight	yellow	Position	6A.01.02 (resolve aircraft conflict situations)
	Airspace conflict	cyan	Position	6A.01.02 (resolve aircraft conflict situations)
ARTS	Owned aircraft	white	Greater luminance	8A.01.01 (perform situation monitoring)
	Unowned aircraft	green	None	8A.01.01 (perform situation monitoring)
	Pointout	yellow	None	8A.01.04.08 (issue pointouts); 8A.01.04.09 (respond to pointouts)
	Alert	red	Flashing alert text right of datablock	8A.01.02 (resolve aircraft conflict situations)
	Weather 1	dark gray	None	8A.01.05 (assess weather impact)
	Weather 2	brown	None	8A.01.05 (assess weather impact)
	Weather 3	reddish brown	None	8A.01.05 (assess weather impact)
STARS	Owned aircraft	white	Greater luminance	8A.01.01 (perform situation monitoring)
	Unowned aircraft	green	None	8A.01.01 (perform situation monitoring)
	Pointout	yellow	Flashing datablock text, with "PO" to the right of the callsign	8A.01.04.08 (issue pointouts); 8A.01.04.09 (respond to pointouts)
	Alert	red	Flashing alert text above datablock	8A.01.02 (resolve aircraft conflict situations)
	Weather 1	dark gray blue	None	8A.01.05 (assess weather impact)
	Weather 2	dark mustard	None	8A.01.05 (assess weather impact)